Quantitative Estimation from Multiple Cues: Test and Application of a New Cognitive Model

D i s s e r t a t i o n

zur Erlangung des akademischen Grades
Dr. rer. nat. im Fach Psychologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät II
der Humboldt-Universität zu Berlin

von
Dipl. Psych. Bettina von Helversen

Quantitative Estimation from Multiple Cues: Test and Application of a New Cognitive Model

D i s s e r t a t i o n

zur Erlangung des akademischen Grades
Dr. rer. nat. im Fach Psychologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät II
der Humboldt-Universität zu Berlin

von
Dipl. Psych. Bettina von Helversen,
geboren am 27.12.1977 in Freiburg im Breisgau

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II
Prof. Dr. Wolfgang Coy

Gutachter/Gutachterin

1. Prof. Gerd Gigerenzer
2. Prof. Peter Frensch
3. Prof. Peter Juslin

Tag der Verteidigung: 18.01.2008

Acknowledgements

This dissertation is the result of research I have carried out at the Center for Adaptive Behavior and Cognition (ABC) of the Max Planck Institute for Human Development (MPI) as a fellow of the LIFE Research school.

First and foremost I would like to dearly thank my advisor Jörg Rieskamp for his time, unfailing support and outstanding mentoring throughout the last three years. I deeply appreciate that he was always there when I needed advice. Jörg's unique ability to reduce a complex problem to a simple question and to combine scientific and methodological rigor with a pragmatic approach, encouraged me to search for simple solutions. Further, I would like to thank my colleagues and friends at the ABC Research group and in the LIFE Research School who with their warm atmosphere, critical minds, and helpful advice made my dissertation possible, and the last three years an exciting time. I'm deeply thankful to Gerd Gigerenzer for providing me the opportunity to work in this wonderful environment and Paul Baltes, and the LIFE directors for giving me the chance to enjoy the truly stimulating experience of being part of the LIFE program.

I'm very grateful to Peter Juslin, Linnea Karlsson, and Henrik Olsson for generously sharing their data with me, enabling me to include a reanalysis of their study in my dissertation. Further I'd like to thank Lael Schooler, Henrik Olsson, Konstantinos Katsikopoulos, Stefan Krauss, Peter Frensch, and Richard Gonzalez for their advice and comments on prior versions of the manuscripts. I also thank Jing Qian, Benjamin Scheibehenne, Rui Mata, Thorsten Pachur, Andreas Wilke, Wolfgang Gaißmeier, and Tim Johnson for their thought-provoking comments, interesting discussions, and most of all their friendship. Above all I want to thank Jutta Mata, for her friendship, patience, encouragement, invaluable help and advice on countless occasions throughout the last three years.

Further I would like to thank Gregor Caregnato and all the student assistants for their help and patience in collecting my data and developing the experimental programs; Christian Elsner for his technical as well as emotional support; Uwe Czienkowski for encouraging me to program the experiments myself and his help with any problems I encountered on the way and Anita Todd for her thorough editing of my manuscripts.

I thank my family and friends who gave me the strength to go forward and helped me find time to relax and enjoy life while working on my dissertation; my Nonna Ruth for her

unfaltering belief in me; my brothers Thomas and Martin for being there, whenever I need them and my parents for their unconditional support. Their dedication to science, insatiable curiosity and questioning minds raised my interest in science and encouraged me to pursue an academic career.

English Summary

How do people make quantitative estimations, such as estimating a car's selling price? Often people rely on cues, information that is probabilistically related to the quantity they are estimating. For instance, to estimate the selling price of a car they could use information, such as the car's manufacturer, age, mileage, or general condition. Traditionally, linear regression type models have been employed to capture the estimation process. These models assume that people weight and integrate all information available to estimate a criterion. In my dissertation, I propose an alternative cognitive theory for quantitative estimation: The mapping model, inspired by the work of Brown and Siegler (1993) on metrics and mappings, offers a heuristic approach to decision making. In the first part of my dissertation, I laid the theoretical foundation for the mapping model, and tested this against established alternative models of estimation, namely, linear regression, an exemplar model, and a simple estimation heuristic. The mapping model provided a valid account of people's estimates outperforming the other models in a variety of conditions. Consistent with the "adaptive toolbox" approach on decision making (Gigerenzer & Todd, 1999), which model was best in predicting participants' estimations was a function of the task environment. In the second part of my dissertation, I further investigated how task characteristics influence the models' ability to predict participants' estimations by focusing on the assumptions the models make about the estimation process: While the exemplar model relies on the establishment of an exemplar memory base, the mapping model requires the abstraction of knowledge. I examined how different task features affect these assumptions and thus explain shifts in processing contingent on the task structure. My results indicate that explicit knowledge about the cues is decisive. When knowledge about the cues was available, the mapping model was the best model; however, if knowledge about the task was difficult to abstract, participants' estimations were best described by the exemplar model. In the third part of my dissertation, I applied the mapping model in the field of legal decision making. In an analysis of fining and incarceration decisions, I showed that the prosecutions' sentence recommendations were better captured by the mapping model than by legal policy modeled with a linear regression. These results indicated that the mapping model is a valid model which can be applied to model actual estimation processes outside of the laboratory. Furthermore, they suggest that deviations from legal policy can be explained by considering the cognitive processes of the decision maker.

Deutsche Zusammenfassung

Wie schätzen Menschen quantitative Größen wie zum Beispiel den Verkaufspreis eines Autos? Oft benutzen Menschen zur Lösung von Schätzproblemen sogenannte Cues, Informationen, die probabilistisch mit dem zu schätzenden Kriterium verknüpft sind. Um den Verkaufspreis eines Autos zu schätzen, könnte man zum Beispiel Informationen über das Baujahr, die Automarke, oder den Kilometerstand des Autos verwenden. Um menschliche Schätzprozesse zu beschreiben, werden häufig linear additive Modelle herangezogen. Diese Modelle nehmen an, dass Menschen alle Informationen, die sie zur Verfügung haben, gewichten und dann zu einer Schätzung integrieren, indem sie die gewichteten Informationen addieren. In meiner Dissertation schlage ich ein alternatives Modell zur Schätzung quantitativer Größen vor. Das Mapping-Modell präsentiert einen heuristischen Ansatz auf der theoretischen Grundlage von Brown und Sieglers (1993) Arbeit zu *metrics* und *mapping*s. Im ersten Kapitel meiner Dissertation lege ich die theoretische Basis des Mapping-Modells dar und teste es gegen weitere, in der Literatur etablierte, Schätzmodelle wie zum Beispiel eine lineare Regression, ein Exemplar-Modell und eine Schätzheuristik. Es zeigte sich, dass das Mapping-Modell unter unterschiedlichen Bedingungen in der Lage war, die Schätzungen der Untersuchungsteilnehmer akkurat vorherzusagen. Allerdings bestimmte die Struktur der Aufgabe — im Einklang mit dem Ansatz der „adaptiven Werkzeugkiste“(Gigerenzer & Todd, 1999) — im großen Maße, welches Modell am besten geeignet war, die Schätzungen zu erfassen. Im zweiten Kapitel meiner Dissertation greife ich diesen Ansatz auf und untersuche, in wie weit das Zusammenspiel von Aufgabenstruktur und den Annahmen, die die Modelle zum Schätzprozess machen, bestimmt, welches Modell die Schätzprozesse am Besten beschreibt. Das Exemplar-Modell setzt die Speicherung von Exemplaren im Gedächtnis voraus, während das Mapping-Modell die Abstraktion von explizitem Wissen über die Aufgabe postuliert. Meine Ergebnisse zeigten, dass die Struktur der Aufgabe beeinflusste, welches Modell die kognitiven Prozesse am Besten beschrieb. Das Mapping-Modell war am Besten dazu geeignet die Schätzungen der Versuchsteilnehmer zu beschreiben, wenn explizites Wissen über die Aufgabe vorhanden war, während das Exemplar-Modell den Schätzprozess erfasste, wenn die Abstraktion von Wissen schwierig war. Im dritten Kapitel meiner Dissertation, wende ich das Mapping-Modell auf juristische Entscheidungen an. Eine Analyse von Strafakten ergab, dass das Mapping-Modell Strafzumessungsvorschläge von Staatsanwälten besser vorhersagte als eine lineare Regression. Dies zeigt, dass das Mapping-

Modell auch außerhalb von Forschungslaboratorien dazu geeignet ist menschliche Schätzprozesse zu beschreiben. Weiter weisen die Ergebnisse darauf hin, dass Abweichungen von gesetzlichen Regelungen auf die kognitiven Prozesse der Entscheidungsträger zurückgeführt werden können.

TABLE OF CONTENT

Introduction

Quantitative estimation is an important task that people have to master in their daily lives, such as estimating the travel time for a journey, the risk of medical treatment, or the quality of a job applicant. For this task people rely on several diverse mechanisms. For instance, numerical estimates can be directly retrieved from memory, reconstructed, for example, from landmark dates in temporal estimation (Friedman, 1993, 2004) or estimated from rates of behavioral frequency (Conrad, Brown, & Cashman, 1998). In my dissertation, I focus on a further mechanism of quantitative estimation: estimation from probabilistic information.

To estimate a quantity of interest, people can rely on multiple sources of information, for instance, cues, which are probabilistically related to the criterion, that is, the quantity being estimated. For instance, to estimate the selling price of a house, people could rely on information, such as the house size, the quality of the neighborhood, or if it has a swimming pool. A variety of cognitive models has been proposed to describe the cognitive processes involved in quantitative estimations, with the purpose to clarify which information people rely on and how they use and integrate multiple pieces of information. Traditionally, linear models, such as multiple linear regressions, have been the model of choice dominating the literature on multiple cue judgments (Hammond & Stewart, 2001; Brehmer, 1994). However, recently, linear regression approaches have been criticized, and the need for more cognitively oriented models postulated (Gigerenzer & Todd, 1999). Since then, several alternative models have been proposed (Juslin, Karlsson, & Olsson, in press; Hertwig, Hoffrage & Martignon, 1999). In my dissertation, I propose a new cognitive theory for estimation from multiple cues, and test it against established models of estimation as well as newly proposed models.

The Traditional Approach to Estimations: Social Judgment Theory

Following the work of Egon Brunswik (1952) and Ken Hammond (1955), multiple linear regression became the dominant model to describe multiple cue judgments (Brehmer & Brehmer, 1988; Brehmer, 1994). Following from this seminal work "social judgment theory" was established (for an overview, see Doherty & Kurz, 1996). According to social judgment theory, human estimation follows a linear additive strategy that can be captured by a regression model (Doherty & Brehmer, 1997). The linear additive approach assumes that

first each cue is weighted according to its importance. Then, an estimate is reached by adding up the weighted cue values (Cooksey, 1996). Optimal cue weights are found analytically by minimizing the squared deviation between the estimated quantity and the estimation (Cohen, Cohen, West & Aiken, 2003).

Since their introduction to judgment research in the 1950s, regression models have been employed to model judgment policies in many domains, reaching from predicting teachers' evaluations (Cooksey, Freebody, & Davidson, 1986), medical decisions (Wigton, 1996), analyzing psychiatrists' diagnostic strategies (Harries & Harries, 2001), or modeling the bailing policies of judges (Ebbesen & Konecni, 1975). Linear additive models have also been very influential in other areas of psychology; prominent examples include, among others, Anderson's (1981) "information integration theory," or the work of Fishbein and Ajzen (1980) on the impact of attitudes and social norms on behavior. However, despite the success of linear models in describing the outcome of a cognitive process (i.e., the final estimation), they have been criticized for not capturing the process itself (Brehmer, 1994; Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Hoffman, 1960; for a review, see Doherty & Brehmer, 1997).

The Exemplar-Based Approach to Estimation

Recently, exemplar models have been suggested as alternative models for explaining human estimation processes. Exemplar models have been successful in modeling the cognitive process underlying categorizations (Nosofsky & Johansson, 2000; Kruschke, 1992). Due to this success, they recently have been considered as models of estimations (Juslin, Olsson, & Olsson, 2003; Juslin et al., in press). Exemplar models assume that encountered objects are stored in memory and retrieved if a new object is evaluated. The estimation is based on a judgment of similarity between the object under evaluation and the exemplars stored in memory. For instance, a professor evaluating the success of a prospective graduate student might think about former graduate students and estimate the success of the prospective student based on the similarity to the former students.

The more similar an exemplar is to the object under evaluation, the stronger its impact on the estimation. The final estimate is given by the average of the criterion values of all stored exemplars, weighted by their similarities to the object under evaluation. Similarity is conceptualized as cue or feature based, that is, objects are described by their values on a list of features. Two objects are considered similar if their values on the features match,

however, the features can differ in their importance for the similarity evaluation, for instance, two objects matching on all but one feature can still be considered as different if this cue is of central importance. Similarly, a mismatch on a feature can be negligible if this feature is of minor importance. The overall evaluation of similarity is reached by integrating all features based on the context model (Medin & Schaffer, 1978).

Juslin et al. (in press) argued that people's estimation processes can be best captured by exemplar models in nonlinear environments, that is, when the cues are nonlinearly connected with the criterion. Consistent with this argument, they showed in several experiments that the exemplar model was better suited than the linear additive rule in predicting estimations when the criterion was a multiplicative function of the cues. Thus, the exemplar model seems to be a valid model for quantitative estimation.

Heuristic Approach to Estimations

A further recent approach to decision making comes from the literature on heuristics. In decision tasks, such as in paired comparison tasks, simple heuristics like Take The Best (Gigerenzer & Goldstein, 1996) have successfully been employed to model the decision process. Especially in complex decision situations and under time pressure, the simple heuristics were better suited to describe behavior than more complicated models based on optimization procedures (Rieskamp, 2006; Rieskamp & Hoffrage, in press; Rieskamp & Otto, 2006; Bröder, 2000; Bröder & Schiffer, 2003). This research indicates that, in many real-life situations, simple heuristics can predict human behavior well. In a similar vein, Hertwig et al. (1999) proposed a heuristic for estimation, QuickEst. QuickEst is a noncompensatory model, that is, it does not integrate information, but bases its estimation on only one cue. The cue on which the estimation is based is found by sequentially searching through all available cues. Once a cue fulfils a previously set criterion, search is stopped and an estimation is made. Although QuickEst makes accurate estimations in environments with a skewed criterion distribution, so far there is no evidence that it can model human estimation processes (Hausmann, Läge, Pohl, & Bröder, 2007). This leaves the question, if estimation processes can be modeled by simple heuristics, how can a heuristic model of quantitative estimations be devised, and in which conditions can it describe human behavior?

A New Cognitive Theory for Quantitative Estimations from Multiple Cues: The Mapping Model

The goal of my dissertation work was to develop a simple cognitive theory that would capture the cognitive process underlying quantitative estimations. Inspired by the research of Brown and Siegler (1993), I developed the mapping model and tested it against established models of estimation. Although Brown and Siegler provide a comprehensive framework of quantitative estimation, they do not offer a computational model of the estimation process. Thus, the goal was to develop a computational model that is consistent with the framework of Brown and Siegler.

The Mapping Model

Brown and Siegler postulated that two types of knowledge are necessary to make an estimation. First, knowledge about the *mapping* properties of the objects is required. This knowledge reflects the ordinal relation among objects, that is, how high an object ranks on the criterion of interest, compared to the other objects. Second, knowledge about the *metric* properties of the criterion is necessary, such as the distribution, the mean, or the range. The mapping model describes how knowledge about the mapping properties of an object is linked to the metric properties of the criterion in the estimation process. In a first step, the mapping models use the cue information to capture the mapping properties of an object. Objects are grouped together according to their cue sums, inferring the ordinal relations of the objects from the number of positive cue values. Second, to represent the metric properties of the criterion, a typical criterion value is derived for each category by considering the criterion values of the objects falling into the same category.

The mapping model only uses binary cue information so that each cue can have either a positive or a negative value. Cues are coded so that they are positively correlated with the criterion. To group the objects together, the mapping model makes the simplifying assumption that all cues are equally important, thus, all objects that share the same number of positive cue values are put into the same category. In the second step, the mapping model derives a typical criterion value for each of the cue sum categories, represented by the median criterion value of the objects in the same cue sum category. To evaluate a new object, the mapping model computes its cue sum and estimates the typical criterion value corresponding to the cue sum category.

Dissertation Outline

In my dissertation, I propose the mapping model as a new model for quantitative estimation, and test it in several experimental studies against other competing theories of estimation. This dissertation is structured into three chapters that are based on three manuscripts.

The first chapter, *The Mapping Model: A Heuristic for Quantitative Estimation*, focuses on the theoretical foundations for the mapping model. Past research had focused on linear regression as the predominant model to analyze quantitative estimations. However, recently, regression models were criticized because they do not describe the cognitive process underlying estimation (Hoffman, 1960; Gigerenzer & Todd, 1999), and the need for more process-oriented models was voiced (Payne, Bettman, & Johnson, 1993; Gigerenzer & Todd, 1999). In response to this criticism, several alternative models were proposed to capture human estimation (e.g., Juslin et al., 2003). Thus, the goal of the first chapter was to derive a model that can not only capture the outcome of estimations but also provides a plausible description of the cognitive process. Furthermore, the model would need to compete with alternative approaches to quantitative estimation.

The framework for quantitative estimation by Brown & Siegler (1993) presents a plausible and comprehensive account of estimation processes. However, it lacks the precise formulation of a computational approach to quantitative estimation. Thus, the aim was to provide a computational model that can be integrated into the framework of Brown and Siegler (1993), and test it rigorously against current models of estimation: a linear additive model, an exemplar model (Juslin et al., 2003), and the heuristic QuickEst (Hertwig et al., 1999) in varying task environments.

In the second chapter, *Models of Quantitative Estimations: Rule-based and Exemplar-Based Processes Compared,* I focus on a comparison of the exemplar model and the mapping model. In this chapter, I follow up on some open questions in Chapter 1. First, the exemplar model and the mapping model both provide an account for estimation processes in situations in which linear additive strategies are less successful. However, the models assume quite different estimation processes. While the exemplar model proposes an implicit similarity-based process, the mapping model assumes a rule-based estimation process. Thus, one goal of the second chapter was to clarify under which conditions the two models describe human estimation. More specifically, I investigated the role of two cognitive components which are essential for the assumptions that the models make about the

estimation process: exemplar memory and knowledge abstraction. I examined which task features affect these components and thus could be responsible for a shift from rule- to exemplar-based processing. Secondly, in the first chapter, I concentrated on quantitative measures to compare the models. In the second chapter, my goal was to devise and include a qualitative test that would allow for the models' assumption to be tested more directly. By constructing situations where due to the assumptions about the estimation processes they make qualitatively different predictions, I provided a more rigorous test of the models assumptions.

The third chapter, *Predicting Sentencing for Low-Level Crimes: A Cognitive Modeling Approach,* presents an application of the mapping model to a real-world problem. In the previous chapters, the mapping model was exclusively tested on laboratory data. However, if the mapping model aims to provide a plausible account of estimation, it also needs to perform well on real data. Sentencing decisions provide an interesting application, as they are a common estimation problem, but resemble the laboratory tasks in several ways. In sentencing a continuous criterion, the magnitude of the sentence has to be determined on the basis of multiple cues, the characteristics of the offense and the offender. In addition, sentences often have a highly skewed distribution, and thus offer an especially interesting task because the mapping model performed well in a similar environment in the laboratory. Moreover, in the legal domain, a recent discussion has raised the question of, how far can legal decision makers abide the law (Dhami & Ayton, 2001; Gigerenzer, 2006). This question is highly relevant because sentencing decisions provide highly complex material, and are often made under time pressure, making it probable that legal decision makers deviate from the rather complex legal regulations. Here, the mapping model could make a contribution by highlighting the importance of the cognitive process for decision making.

Chapter 1:
The Mapping Model: A Heuristic for Quantitative Estimation

Abstract

How do people make quantitative estimations, such as estimating a car's selling price? Traditionally linear-regression-type models have been employed to answer this question. These models assume that people weight and integrate all information available to estimate a criterion. We propose an alternative cognitive theory for quantitative estimation: The mapping model, inspired by the work of Brown and Siegler (1993) on metrics and mappings, offers a heuristic approach to decision making. We test this model against established alternative models of estimation, namely, linear regression, an exemplar model, and a simple estimation heuristic. With four experimental studies we compare the models under different environmental conditions. The mapping model proved to be a valid model to predict people's estimates.

The Mapping Model: A Heuristic for Quantitative Estimation

Estimating unknown quantities represents a judgment problem encountered frequently in daily life. People estimate the selling price of cars, the productivity of job candidates, or the travel time for journeys. To make these estimates, people use cues that are probabilistically related to the quantity being estimated; for instance, the selling price of a car can be estimated on the basis of the car's mileage, age, or accident record. How do people make estimates? We approach this central question by introducing a new cognitive model—the mapping model. We test this model against alternative models of human estimation.

Beginning with the work of Ken Hammond (1955), who was in turn inspired by Egon Brunswik's ideas (e.g., Brunswik, 1952), linear additive models have been the standard for describing human judgments (Gigerenzer & Kurz, 2001). The research on "social judgment theory" (for an overview, see Doherty & Kurz, 1996) that followed from this seminal work encompasses a large body of studies examining people's judgments in many areas, including, among others, clinical judgments (Harries & Harries, 2001; Wryobeck & Rosenberg, 2005), teachers' evaluations of student achievement (Cooksey, Freebody, & Davidson, 1986), bail decisions (Ebbesen & Konecni, 1975), personnel selection and evaluation (Zedeck & Kafry, 1977), and medical decision making (Wigton, 1996; for reviews see Brehmer & Joyce, 1988; Brehmer, 1994). In all these studies, people's judgments are described by fitting a regression model to the data. Following the tradition of social judgment theory (Hammond, 1996) we hitherto refer to the quantity being estimated as the criterion and to the information used to estimate the criterion as the cues. Like the broader class of linear additive models, linear regression assumes that for each cue, the relation between the cue and the criterion is abstracted and represented by a weight, where the specific weight of a cue defines the cue's impact on the final estimation.

The strong influence of linear additive models is not restricted to research on judgment and decision making. For instance, the linear additive model was employed in Anderson's (1981) "information integration theory," which describes integration of social as well as physical information. Likewise it was adopted to describe the impact of social norms on behavior (Fishbein & Ajzen, 1980). Despite the model's success in describing human behavior, in the present article we challenge the assumption that the underlying cognitive process of human judgment follows the additive integration of weighted information. In its

stead we propose the mapping model as a new model of human estimation. This model is based on Brown and Siegler's (1993) work on metrics and mapping. Our main goal was to test this model rigorously against a linear additive model, and additionally against alternative recent cognitive models of human estimation.

The Mapping Model

Brown and Siegler (1993; see also Brown, 2002) suggested that real-world quantitative estimations rely on knowledge about the *mapping* properties of the objects and the *metric* properties of the criterion. The mapping properties reflect the ordinal relations among the objects in one domain, that is, the knowledge about which object will have a higher value on the criterion compared to other objects. Knowledge about the metric properties, on the other hand, refers to the statistical properties of the criterion, such as the mean, the median, and the functional form of the distribution. Brown and Siegler (1993) assumed that to make accurate quantitative estimations, knowledge about both types of properties is indispensable, yet they did not specify a computational model describing human estimation. Therefore we suggest one that is inspired by the ideas of mapping and metrics.

The mapping model specifies how knowledge about the mapping and metric properties of objects is acquired in two separate steps. First, knowledge about the mapping properties is gathered from the cues. The sum of the cue values is used to infer the ordinal relations of the objects and to group them into categories. Second, to represent the metric properties of the criterion, a typical criterion value is derived for each category by considering the criterion values of other objects falling into the same category. The mapping model only uses binary cue information, so that each cue can have either a positive or a negative value. Cues are coded so that they are positively correlated with the criterion. The knowledge about the mapping properties is then derived by a simple counting strategy, adding up the positive cue values for all cues J of each object i and categorizing them according to their cue sums:

$$(1)\quad k_i = \sum_{j=1}^{J} c_{ji}$$

where k denotes the cue sum of object i and c_{ji} refers to the cue value of object i on cue j.

For each cue sum category a typical criterion value is abstracted, represented by the median criterion value of all known objects that share the same cue sum.[1] To estimate the criterion value of a new object, the probe (p), the cue sum of the probe is computed and the typical criterion value of the corresponding cue sum category is used as an estimate:

(2) $\hat{y}_p = Mdn(x_i, k_i = k_p)$,

where $\hat{y}_p$ denotes the estimated criterion value for probe *p*, which is estimated by the median (*Mdn*) of the criterion values of all known objects *i* that belong to the group of objects with the same cue sum *k* as the probe *p*. If a cue sum category does not exist because no object with a corresponding cue sum was encountered in the past, the average value of the adjacent categories is employed as an estimate.

We demonstrate the mechanism of the mapping model with the illustrative example of estimating the selling price (i.e., the criterion) of two mobile phones, let's call them Psi and Omega, offered in an online marketplace. The phones' features (i.e., weight, display size, digital camera, and Internet access) can be employed as cues to estimate the selling price. To estimate the selling prices of Psi and Omega we can compare them on the features to four similar phones, A, B, C, and D, that were sold in the past (see Table 1). The mapping model estimates that phone Psi will sell for $100, because of the four phones sold (A–D), only phone D—which sold for $100—falls into the same cue sum category. For phone Omega with a cue sum of one, the mapping model estimates the median price of the two phones A and B with the same cue sum, which sold for $10 and $20, respectively, yielding an estimated selling price of $15.

Table 1: Mobile Phone Example for Illustrating the Predictions of the Models

	Phone A	Phone B	Phone C	Phone D	Phone Psi	Phone Omega
			Cues			
Digital camera	-	-	-	+	+	+
Internet access	-	+	+	-	+	-
Weight	-	-	+	+	+	-
Display size	+	-	-	+	-	-
Criterion (selling price, in dollars)	10	20	30	100	?	?
			Estimations of the models (in dollars)			
Mapping	15	15	30	100	100	15
Regression	10	20	30	100	110	90
QuickEst	15	15	20	50	30	15
Exemplar	10	20	30	100	30	43

Note. A plus sign indicates a positive cue value—for example, the phone possesses a digital camera or is lightweight; A minus sign indicates a negative cue value—for example, the phone does not possess a digital camera or it is heavy. Question marks indicate that the selling prize is unknown.

Alternative Theories of Estimation

With the mapping model, we question the widespread assumption in cognitive psychology that human judgments follow a linear additive process of information integration. We first test the mapping model against the most established representative of linear additive models—linear regression. Because other models have recently been proposed to explain estimations from multiple cues, the mapping model is also tested against two of these competitors: an exemplar model (Juslin, Olsson, & Olsson, 2003b) and a heuristic strategy (Hertwig, Hoffrage, & Martignon, 1999). We use our illustrative example to explain the models and show how their predictions differ.

Multiple linear regression. Linear additive models assume that explicit cue–criterion relationships are abstracted and represented as cue weights. Multiple linear regression (MLR) computes optimal weights for every cue, minimizing the squared deviations of the prediction from the criterion (e.g., Cohen & Cohen, 2003). The weights indicate how much impact a given cue has on the estimate of the criterion. The estimated criterion value, $\hat{y}_p$, of the probe p is given by the sum of the product of the cue values, c_j, of the cues j with their respective weights, ω_j, plus an intercept, ω_0:

$$(3)\quad \hat{y}_p = \sum_{j=1}^{J} \omega_j c_j + \omega_0$$

In our example, the four sold phones are used to fit the regression model. That is, the model finds the weights that minimize the squared deviation of the predicted from the real criterion value of the phones sold. In our example optimal weights for the cues are 80, 10, 10, and 0, respectively, with an intercept of 10. The fitted regression model then predicts a selling price of \$110 and \$90 for the new phones Psi and Omega, respectively.

In addition we tested two simplified versions of this standard regression model. First, we included a stepwise regression model that includes only significant parameters (Hastie, Tibshirani, & Friedman, 2001). Second, we tested a simplified version of the regression model that was not fit to participants' estimations. Instead, the optimal parameters for solving the task were selected a priori based on the objective criterion values. However, across all of the following studies the standard regression model was most successful in predicting participants' estimations for new independent observations that were not used to estimate the models' parameters, so that for the sake of clarity we only report the results for the standard regression model.

Exemplar-based model. A promising alternative approach to quantitative estimation is provided by exemplar-based models (EBMs), which in the past have been successfully applied to explain human categorization (for an overview see, for example, Nosofsky & Johansen, 2000). Exemplar models assume that people categorize objects by determining how similar they are to formerly encountered exemplars of the categories and assigning them to the category with the most similar exemplars. Thus, in contrast to a linear additive model, exemplar models do not assume the abstraction of cue–criterion relationships but rely on a knowledge base of memorized exemplars. Recently, Juslin et al. (2003b; Juslin, Jones, Olsson, & Winman, 2003a) reformulated the original context model of Medin and Schaffer (1978) for the area of quantitative estimation (see also Dougherty, Gettys, & Ogden, 1999; Juslin & Persson, 2002; Smith & Zárate, 1992). Juslin, Karlsson, and Olsson (in press, see also Olsson, Enqvist, & Juslin, 2006) showed that exemplar models are more suitable for predicting people's estimations than linear regression when the cues are nonlinearly related to the criterion.

The exemplar model proposed by Juslin et al. (2003a, b) is closely related to the generalized context model[2] (Nosofsky, 1986, 1992; Nosofsky & Johansen, 2000). Exemplar models assume a memory-based inference process. To estimate the criterion of a new object (the probe), the similarity of the probe to the exemplars retrieved from memory is determined. The more similar the probe is to an exemplar, the closer the estimate will be to the exemplar's criterion value. The final estimate of the criterion is the average of the criterion values of all memorized exemplars, weighted by their similarities to the probe:

$$(4)\quad \hat{y}_p = \frac{\sum_{i=1}^{I} S(p,i)\cdot x_i}{\sum_{i=1}^{I} S(p,i)},$$

where $\hat{y}_p$ is the estimated criterion value for probe p; S is the similarity of the probe p to the stored exemplars i with the criterion value x_i; and I is the number of stored exemplars in memory. The similarity S between the probe and an exemplar is determined by the multiplicative similarity rule of the context model (cf., Medin & Schaffer, 1978):

$$(5)\quad S(p,i) = \prod_{j=1}^{J} d_j ,$$

where the variable d specifies the similarity between the probe and the exemplar on the cue dimension j, and d_j takes the value 1 if the values of the probe and the exemplar on cue

dimension *j* match and s_j if they do not. The parameter s_j is an attention weight parameter capturing a cue's importance for the similarity assessment and varies between 0 and 1. A large value for the attention parameter *s* close to 1 implies that a mismatch on this cue has almost no effect on the overall similarity, whereas a low value for *s* close to 0 implies that the cue is very important, because the overall similarity approaches zero if the cue values do not match.

The standard exemplar model assumes that the importance given to each cue varies by using different attention parameters (e.g., Juslin et al. 2003a, b). However, by having one free parameter for each cue the exemplar model is relatively complex and it is an open question whether this complexity is required to capture the underlying cognitive process of estimations. To answer this question we additionally implemented a simplified version of the exemplar model, which assumes that only one single attention parameter *s* is used for all cues (see also Juslin & Persson, 2002). This single parameter then represents the gradient of the similarity function; that is, if *s* is close to 0 only very similar exemplars will influence the estimation, but if *s* is close to 1 also less similar exemplars will be considered. Finally, we implemented a third version of the exemplar model that did not fit parameters to participants' estimations; instead, the parameter values were derived by using the objective criterion values of the objects in the training phase. It turned out that the simplified exemplar model with only one free parameter was most successful in predicting individuals' estimations for new independent observations, so that for the sake of clarity we only report the results for the simplified exemplar model with the exception of the following simulation study and Study 4. When the simplified exemplar model is applied to our phone example, using an attention parameter of $s = .001$ to predict the phones' selling prices, the selling prices of phone Psi and Omega were estimated to be \$20 and \$43, respectively.

A heuristic for estimation—QuickEst. Although regressions models are able to describe the outcome of a cognitive process (i.e., the final estimation), they have been criticized for not capturing the process itself (Brehmer, 1994; Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Hoffman, 1960; for a review see Doherty & Brehmer, 1997). Gigerenzer, Todd, and the ABC Research Group (1999) have argued that the cognitive process of making judgments can often be best described with simple heuristics. Recent experimental work has illustrated that simple heuristics can predict people's inferential choices well, in particular when the application of complex strategies is more costly (e.g., Bröder, 2000; Bröder & Schiffer, 2003; Rieskamp, 2006; Rieskamp & Hoffrage, in press; Rieskamp & Otto, 2006). In this

vein, Hertwig et al. (1999) proposed a heuristic for quantitative estimations, QuickEst, that uses only a small amount of information. According to the heuristic, people process cues sequentially and stop searching as soon as a cue has a negative cue value. Hertwig et al. showed that QuickEst's predictions are as accurate as those of linear regression when applied in an environment where the distribution of the objects' criterion values is J-shaped. A distribution is called J-shaped if most values are small and only a few high values exist, such as, for instance, the distribution of incomes.

QuickEst uses only binary cue information. Each cue can have either a positive or a negative value. All cues are coded such that they correlate positively with the criterion. Accordingly, for each cue, objects with a positive cue value will on average have higher criterion values than objects with a negative value. Next, for each cue the mean criterion value of all objects that have a negative cue value is computed, here called the *nil mean size* (Hertwig, Hoffrage, & Sparr, 2007). Likewise the mean criterion values of the objects with a positive cue value are determined (conditional positive mean). The idea of QuickEst is to stop searching for more information as soon as it becomes probable that an object has a small criterion value. Thus QuickEst stops search as soon as a cue with a negative cue value is encountered or if the cue value for the object is missing. If a positive cue value is encountered, the next cue is considered until all relevant cues have been looked up. QuickEst searches through the cues according to their nil mean size beginning with the smallest.

In contrast to Hertwig et al. (1999) we assume that the maximum number of cues that are searched for is a free parameter capturing individual differences. An estimation is based on the cue that stopped search. If the search was stopped because a negative cue value was encountered, the nil mean size of that cue is used as an estimate. If search was stopped because the maximum number of cues had been considered, the conditional positive mean of the last cue is estimated. For the estimates the means are rounded to the next spontaneous number[3] (Albers, 2001). For our phone example, the nil mean sizes of the cues are 20, 55, 15, and 25, respectively. QuickEst starts search by looking up the information of the phones' weight, the cue with the smallest nil mean size. If this cue has a positive value, it continues search, considering whether the phone has a digital camera, and so on. Because phone Psi has a positive value on weight and has a digital camera, search continues until information for display size is looked up. As phone Psi has a negative value on display size, the rounded conditional mean of this cue ($30) is estimated as the selling price. For phone Omega, search

stops after looking up information for weight, and its nil mean size of $15 is estimated as the selling price.

Testing the theories. Conceptually the theories we consider can be distinguished by two aspects: (1) the way they abstract knowledge from objects encountered in the past, that is, their *knowledge abstraction assumptions*; and (2) the way the abstracted knowledge and the information a probe provides is processed to make a final estimation, that is, their *process assumptions*.

The regression model assumes an additive estimation rule. To build this estimation rule it abstracts knowledge about the cue weights from the encountered objects, taking the dependencies between cues into account. Once this rule is established, previously encountered objects can be forgotten. For the estimation process the model integrates all available information, determining a weighted sum of the cue values. Like the regression model, QuickEst assumes that knowledge, that is, the mean criterion values of the cues, is abstracted from encountered objects. However, QuickEst does not integrate any information; instead cues are searched sequentially and an estimation is made on the basis of one single cue. The exemplar model does not abstract much knowledge; instead it assumes that all encountered objects are stored in memory. Nevertheless, the knowledge of how much attention a cue receives is abstracted from the encountered objects. For the estimation process the exemplar model assumes that the information of all stored exemplars is integrated, by determining a mean of the retrieved criterion values weighted by the similarity of the retrieved exemplars to the probe. In sum, the regression model assumes heavy knowledge abstraction from encountered objects and an information integration process for estimation. QuickEst assumes knowledge abstraction and no information integration, and the exemplar model assumes little knowledge abstraction but relies heavily on integration of information for making an estimation.

Similar to QuickEst and regression, the mapping model assumes a rule-based estimation process, relying on the abstraction of knowledge. The mapping model groups objects into categories on the basis of their cue sums, regardless of the pattern of cue values. For each cue sum category the criterion values of the objects falling into this category are stored (see also Footnote 1). For the estimation process the cues' information on the probe is integrated by a simple adding rule. Then for each probe the median criterion value of the corresponding cue sum category is retrieved and used as an estimate.

How does the mapping model compare to the other models? The mapping model resembles QuickEst in the way it abstracts knowledge by categorizing objects into groups. However, while QuickEst bases its estimation on only one cue, the mapping model assumes that the available information is integrated. Similar to regression, the mapping model relies on an additive integration of information. However, it assumes that every cue contributes equally to the cue sum, whereas the regression model assumes differential weighting of cues. Further, the estimation process of the mapping model does not terminate with the integration of the cue values but continues with determining the typical criterion value using the median of the criterion values of the objects falling in the same cue sum category. As in the exemplar model, this retrieval process of the mapping model can be conceptualized as "similarity based," because the retrieval is guided by finding the best match between the cue sum category determined for the probe and the criterion values for the categories abstracted from the objects encountered in the past. However, the exemplar model and the mapping model differ in how they define similarity. The exemplar model assumes that objects are represented in terms of discrete cue values and similarity is a function of the matches and mismatches on each cue. For the mapping model similarity is a strict function of the cue sum category. Thus, although the simplified exemplar model and the mapping model both assume that cues are equally weighted, two objects that the mapping model groups together because they share the same cue sum could be very different for the exemplar model depending on the pattern of cue values.

Although the theories that we consider differ conceptually, empirically they often lead to similar predictions. To test the theories against one another it is therefore important to identify conditions under which the predictions differ. One aspect of the environment has already been shown to differentiate the theories: the distribution of the criterion values. Hertwig et al. (2007) found that QuickEst outperformed linear regression if the criterion distribution was J-shaped but performed poorly when the criterion was uniformly distributed. In J-shaped distributions characteristically only a few objects have high criterion values, while most have low values. Such distributions are so named because they resemble a J (rotated 90 degrees clockwise) if the objects are ordered according to their ranks. Formally, they can often be described by a power function (i.e., $y = b \times x^a$). A distribution following a power law additionally implies that the rank of an object is specifically related to its size, so that if log rank is drawn against log size, a straight line results. Likewise we will

refer to a uniform distribution as a linear distribution, because a straight line results if rank is plotted against size.

The use of a criterion that follows a power function has a further advantage. Test situations that allow discrimination between models often consist of highly artificial cases that are no longer representative of the original problem. Power law distributions, on the other hand, are among the most prevalent distributions encountered in everyday life. Since power law distributions are related to general growth processes (Gabaix, 1999), they can well describe phenomena as diverse as people's incomes, magnitudes of earthquakes, sales of books or music, or the sizes of computer files, moon craters, or cities (Levy & Solomon, 1997; for a review see Schroeder, 1991). Therefore we extend the work of Hertwig et al. (2007) by conducting a simulation study to investigate how all the models discussed here, especially the mapping model, perform in an environment with a J-shaped and a linearly distributed criterion, respectively.

Simulation study

The goal of the simulation study was to examine how accurate the various models are in solving estimation problems under different environmental conditions. Furthermore the goal was to identify environments in which the models make distinct prediction that allow an experimental test.

The simulations were designed to resemble an experimental condition as closely as possible, while still providing enough data to result in reliable evaluations of the models' accuracies. First, J-shaped and linearly distributed criterion values, ranging between 2 and 100, were created for 50 objects by using a power function ($y = bx^a$, with $a = -1$, $b = 100$, and x ranging between 1 and 50) for the J-shaped environment and a linear function ($y = bx + c$, with $b = -2$ and $c = 102$) for the linear environment. To investigate if potential accuracy differences would hold over a wide range of situations, we varied two further factors: The cue–criterion correlation and the percentage of positive and negative cue values per cue (for details see Appendix A).

We examined models' accuracies by cross-validation (averaged over 100 trials). That is, we randomly selected 100 times one half of the data—the calibration sample— to estimate the models' parameters, and then we tested the models on the other half—the validation sample—to test the models' accuracies for new objects. Models' accuracies were

characterized by the root mean square error (*RMSD*) of the models' predictions and the criterion values.

How accurate were the four models? In general, accuracy was strongly affected by the distribution of the criterion value. In the linear environment, the *RMSD* was on average two times larger than in the J-shaped environment. When fitting the data of the calibration sample, all four models performed better than a baseline model, which always predicted the average criterion value of all objects of the calibration sample (see Table 2). The exemplar model was the best model in both conditions, and QuickEst was worst. However, the validation sample represents the crucial situation of making predictions for new objects. Here QuickEst performed best in the J-shaped environment, and the mapping model was second best, $t(31) = 2.15$, $p = .02$, with an effect size of $d = .45$ (Cohen, 1988). In the linear environment, the mapping model was the best in the validation sample, followed by the exemplar model, $t(53) = 3.08$, $p < .01$, $d = .42$, and QuickEst performed worst. These results illustrate that the criterion distribution influences models' accuracies differentially. They are in line with the results of Hertwig et al. (2007), who reported that the accuracy of linear regression is affected negatively by a skewed distribution, whereas the accuracy of QuickEst deteriorates if the criterion is linearly distributed.

Table 2: Models' Average Accuracies (Root Mean Square Error) in the Simulation Study for the Two Environments

Model	J-shaped				Linear			
	Calibration sample		Validation sample		Calibration sample		Validation sample	
	M	SD	M	SD	M	SD	M	SD
Mapping	14.3	3.5	15.3	1.6	21.6	5.1	25.9	6.4
Regression	14	2.4	16.5	1.2	20.9	4.7	27.7	6.3
QuickEst	14.8	1.7	14.9	1.1	24.8	3.5	28.3	3.5
Exemplar	12	3.5	15.8	1.7	17.5	4.9	27.2	6.2

Note. The models were initially fitted to the calibration sample, which contained 50% of the objects; the validation sample was used to cross-validate the results and comprised the other 50% of objects. Model predictions in the validation sample were made by using the parameter values derived in the calibration sample. The variation in model accuracy was higher in the linear environment, as the design in the linear environment varied over a higher number of correlations, and magnitude of correlation affected the accuracy of the models.

The difference in model accuracy between the calibration sample and the validation sample highlights the problem of over-fitting: Complex models with several free parameters

are highly flexible in fitting any data, running the risk of fitting noise instead of fitting systematic structure (see Olsson, Wennerholm, & Lyxzén, 2004; Pitt, Myung, & Zhang, 2002). For this reason in our experimental studies we tested the models by using a generalization test (cf., Busemeyer & Wang, 2000): First, participants made estimations for a training set, which was later used to estimate the models' parameters. Then they made estimations in a test set, which was used to test the models' predictions against each other.

Study 1

Study 1 was designed to test how well the four models of quantitative estimation can predict human estimations. To control for prior knowledge, participants were presented with an artificial inference problem. Following the work of Juslin et al. (2003a, b), participants had to estimate the toxicity of fictional bugs, which were described by five dichotomous cues. For a rigorous test of the models, the experiment varied the distribution of the criterion values in a between-subjects design. In the first condition, the linear environment, the criterion values were linearly distributed, whereas in the second condition, the J-shaped environment, the distribution of the criterion values followed a power law function.

Method

Participants. Sixty participants took part in the experiment: 30 women and 30 men. The participants were randomly assigned to the two experimental conditions, balanced for gender. They were on average 25 years old and most were students from one of the Berlin universities. The data of one participant in the linear environment was later excluded because the participant did not put any effort into solving the task, responding with the same number as an estimate in every trial. Participants were paid according to their performance in the task; the average payment was €13 for an individual session lasting on average 1.5 hr (with €1 corresponding to $1.28 at the time of the study).

Procedure and materials. The study was conducted as a computer-based experiment. Written instructions informed the participants that their task was to estimate the toxicity of different bugs on the basis of five binary cues (color of head, length of antennae, color of wings, size, and biotope). The toxicity of the bugs was measured by the amount of venom in the saliva and could vary between 20 and 1,000 mg per liter. As a cover story the participants were told that the toxicity of the bugs differed depending on the subspecies the bugs belonged to and that the cues would help them to estimate the bugs' toxicity correctly.

The bugs could not be distinguished solely on the basis of the cues, as some of the subspecies were very similar in appearance. In these cases only a genetic test could identify the correct subspecies. To speed up learning of the task, the participants were informed about the direction of the cues, that is, which cue values indicated higher levels of toxicity, without learning the magnitude of the correlation.

Depending on the experimental condition the criterion was either J-shaped or linearly distributed. In both conditions, the experiment consisted of a learning phase, in which the participants could learn to estimate the bugs' toxicity, and a test phase, in which the toxicity of new bugs had to be estimated. In the training phase the participants had to estimate the toxicity of 20 bugs. This phase consisted of 200 trials structured in 10 blocks, each presenting the 20 bugs from the training set in random order. The participants were not told that the same bugs would be repeated; instead each time a bug reappeared, it had a new number. In each trial one bug was presented with its five cue values on the screen and participants were asked to give an estimate of the toxicity of the bug. The order in which the cues were presented was randomly determined for each participant.

After making the estimation, participants were given feedback about the accuracy of their estimate and received points accordingly. Participants' payment was contingent on their performance. After the experiment the total number of earned points was exchanged into euros at a rate of €0.1 for 100 points. For each estimation that exactly matched the correct criterion value, the participants were awarded 100 points. Deviations from the correct criterion value led to fewer points, with increasing inaccuracy leading to a disproportionately larger decrease in points. Specifically, the feedback algorithm used the mean squared deviation of the estimation from the actual criterion value to determine how many points were subtracted from the maximum 100 points for an exact estimation.

To create a moderately exacting feedback environment (Hogarth, Gibbs, McKenzie, & Marquis, 1991), which has been shown to lead to high performance (Gonzalez-Vallejo & Bonham, in press), the feedback algorithm incorporated a correction term to account for the difficulty of the task (see Appendix B for details). The correction term consisted of a constant that determined the magnitude of the deviation that would result in a payoff of zero points. Any deviation exceeding the deviation by the correction term would lead to the subtraction of points. The correction term was chosen so that reliance on a baseline model that always estimated the same value would result in zero points. Since the baseline model reached a better fit in the J-shaped environment, the correction terms in the two

environments differed. In both conditions participants received 100 points for a correct answer; in the J-shaped environment a maximum of 355 points was subtracted for an error whereas a maximum of only 127 points was subtracted in the linear environment. In the instructions it was explained to the participants that subtracting points for errors was employed to correct for chance performance.

In addition to earning points, the participants received outcome feedback on each bug's actual criterion value, the mean squared error of their estimation, and their current total score. In the test phase the participants made the same judgments as in the training phase, but without outcome feedback. They were informed that nevertheless they would earn points according to their accuracy. The test set consisted of 21 profiles that included the old profiles from the training set as well as new profiles.

The training and the test set were constructed so that the models' predictions for the test set, given the training set, would be as different as possible.[4] To find a training set–test set combination that would allow for good discrimination between the models in both environments, we first chose an environment from the simulation in which the models had differed in their predictive accuracy. In this environment each cue had 50% negative cue values and correlated positively with both criteria. We randomly selected 100 training sets of 20 bugs from this environment under the constraint that the highest and the lowest criterion value were always included, ensuring the full range of the criterion for the estimations. All criterion values were multiplied by 10 to have a larger range. Then each model was fitted to the bugs of the training sets, maximizing the model's accuracy in estimating the bugs' toxicity. After fitting the models' parameters, the models' predictions were determined for all objects that did not appear in the training set.

From the 100 training sets we selected the one that allowed the best discrimination between all four models on the new objects, given two additional restrictions. First, to avoid the objection that the participants simply learned to make estimations according to the best performing model in the training set, we excluded all training sets in which the models' accuracy differed widely in the J-shaped environment. Second, we excluded all training sets in which the same cue profile appeared more than four times, to ensure that the differences in model predictions were not due to an extreme training set. Finally from the remaining training sets the one that maximized the differentiability of the models in the test set was selected, which was the set with the highest number of cue profiles for which two models made predictions differing by more than 100 mg/l of estimated toxicity.

The final training set consisted of 20 objects with 20 different criterion values, but with only eight different cue profiles, so that one profile appeared once, three profiles twice, three profiles three times, and one profile four times (see Table 3).

Table 3: Task Structure of Study 1

Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	J- shaped criterion	Linear criterion
0	0	0	0	0	20	20
0	0	0	1	0	23	60
0	0	0	0	0	26	80
0	1	0	1	0	28	140
0	0	0	1	0	29	160
0	0	1	0	1	33	220
0	1	0	1	0	34	240
0	0	1	0	1	35	260
0	1	0	0	0	40	300
0	1	0	1	1	41	420
0	0	0	1	0	47	440
0	1	1	0	1	52	480
0	1	0	1	0	62	540
0	1	0	1	1	71	640
0	1	0	0	0	110	660
0	1	0	1	1	160	720
1	1	1	1	1	200	840
0	1	0	1	1	250	880
1	1	1	1	1	500	920
1	1	1	1	1	1,000	1,000

All cues correlated positively with the criterion and the cue–criterion correlations differed between .30 and .79 (see Table 4). For the test set, the cue profiles for which the four models made the most different predictions were selected. For any pairwise model comparison, at least four profiles allowed a good differentiation between the two models. Also the bugs of the training set were included in the test set. The test sets of the linear and the J-shaped environments and the models' predictions based on the training set can be found in Appendix C.

Table 4: Correlations Between Cues and Criteria in Study 1

	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5
J-shaped criterion	.79	.35	.48	.30	.42
Linear criterion	.65	.66	.37	.39	.62

How well could the different models solve the estimation problem in the training phase? In the J-shaped environment the models' predictions, when fitting the parameters to the objective criterion values, deviated from the criterion with a mean root mean square deviation (*RMSD*) of 136 and could explain about 64% of the variance. QuickEst and the mapping model did slightly worse than the other models. Because the training set in both environments consisted of the same cue profiles, the models' accuracies could not be controlled for in the linear environment, but the accuracies did not differ substantially among the regression, exemplar, and mapping model (M = 145, SD = 11). Only QuickEst with an *RMSD* of 183 did clearly worse than the other models. Although the *RMSD* in the linear environment was higher on average, the models could explain more linear variance (average r^2 = .74).

Results

Overall, the mapping model explained the predictions of the participants best in the test phase, if all conditions were considered jointly. However, the distribution of the criterion played an important role. In the J-shaped environment the mapping model was clearly the best model, whereas in the linear environment the standard regression model and the exemplar model with only one parameter performed equally well. Before we come to the model comparisons, we first report participants' accuracy.

Accuracy of the participants. Participants' accuracy was measured by the *RMSD* between participants' estimations and the criterion and by the Pearson correlation of the estimations with the criterion. Participants were quite successful in learning the bugs' toxicity levels during the training phase, in particular when considering that due to the indistinguishable cue profiles perfection was not possible. The strongest learning effects were observed between the first and the fourth block. Overall, the mean *RMSD* dropped in both environments from 236 (J-shaped) and 232 (linear) in the first block to 149 and 194,

respectively, in the 10th block. The last three blocks showed no significant learning effects, so the data were merged for the further analyses. The average accuracy in the linear environment (*RMSD* = 210) was worse than in the J-shaped environment (*RMSD* = 163), $U = 104$, $p < .01$. However, the average amount of variance explained did not differ; $r^2_{\text{linear}} = .58$, $r^2_{\text{J-shaped}} = .58$.

Estimating the models' parameters. As the primary measure of the models' goodness-of-fit, the *RMSD* between the participants' estimations and the models' predictions was used. The models' parameters were estimated by minimizing the *RMSD* for participants' estimations in the last three blocks of the training phase. The models were tested against each other on the basis of the *RMSDs* of their estimations for the test phase. Additionally we considered the degree of linear variance explained by the models (the coefficient of determination r^2), because the two measures capture slightly different aspects of model fit and r^2 is the preferred measure in the social judgment theory literature. But since the two measures are not independent all model tests are solely based on the models' *RMSD*.[5]

The models were fitted individually to each participant: For the linear regression the parameters were determined analytically using the cues of the training set and the individual participants' estimates. The exemplar model was fitted on the last three blocks of the training phase with the correct cue and criterion values of the training set as the memory base. The best parameter for each participant was searched for by using the quasi-Newton optimization method as implemented in MATLAB. To avoid local minima, parameters were first derived by a grid search with the results serving as the starting values for the subsequent fitting procedure. For QuickEst only one parameter had to be estimated specifying the maximum number of cues considered, and here the optimal parameter value was selected by an exhaustive search. If different numbers of cues reached the same fit the lowest number was selected. The mapping model entails no free parameters, so no parameter was estimated; the medians for the different categories the mapping model used were determined on the basis of the objects' criterion values in the training set.

Model comparison—training phase. We first compared each model's fit with the fit of a baseline model in the training phase, which predicted only one single value for all objects encountered; the specific value the baseline model predicted was fitted to the data of the training phase. The baseline model reached an average fit of *RMSD* = 289 in the linear environment and of 225 in the J-shaped environment. Because the baseline model is a rather naïve model of estimation, any of our four models needs to prove first that it can do better by

taking the dependencies of the estimations on the cue profiles into account. For the training phase all four models did better than the baseline model in predicting participants' estimations (see Table 5). To test if one model could explain participants' estimations significantly better than another model we used a non-parametric test (i.e., the Wilcoxon *Z*-test). In describing the data of the training phase, the regression model did best in both environments, followed by the exemplar model (linear: MLR vs. EBM, $Z = -4.67$, $p < .01$; J-shaped: $Z = -4.78$, $p < .01$), explaining more than 80% of the variance in the linear environment and almost 80% in the J-shaped environment (see Table 5). The mapping model and QuickEst did significantly worse than the other two models, particularly in the linear environment. As described above, for clarity we only report the results for the standard regression model with six free parameters and the simplified exemplar model with one free parameter (for the results of the other versions see Appendix D).

Table 5: Models' Average Accuracies in Predicting Participants' Estimations in Study 1

	Linear environment				J-shaped environment			
	Mapping	Regression	QuickEst	Exemplar	Mapping	Regression	QuickEst	Exemplar
				Training set				
RMSD	149	93	168	138	125	98	125	116
SD	35	26	23	62	41	40	40	37
r^2	.75	.89	.69	.81	.77	.77	.76	.76
SD	0.16	0.07	0.10	0.08	0.17	0.17	0.18	0.17
				Test set				
RMSD	158	166	285	161	139	342	166	166
SD	49	56	46	40	93	124	101	70
r^2	.68	.67	.31	.67	.55	.20	.39	.47
SD	0.17	0.17	0.07	0.13	0.23	0.23	0.24	0.17

Note. The number of participants was 29 in the linear environment and 30 in the J-shaped environment. The exemplar model had one free parameter.

However, the models' fit for the training phase is not very meaningful for testing the models against each other: Even though we tried to put the models on more equal footing, they still differed in their complexities, that is, in the number of free parameters and the complexity of their functional form. Thus it is not surprising that the models with greater flexibility—the regression and the exemplar model—did better in fitting the data than the mapping model. Therefore the crucial model comparison test consists of how well the models predict participants' estimations for new independent objects of the test phase. This generalization test goes beyond a pure cross-validation test, because the new objects of the test phase differed from the objects of the training phase.

Model comparison—test phase. The models' predictions for the test phase were determined on the basis of the estimated parameters of the training phase. The baseline model reached a better fit in the test set than in the training set with an average fit of *RMSD* = 180 in the J-shaped environment and *RMSD* = 282 in the linear environment. This is presumably because the new profiles included in the test set had less extreme cue profiles; that is, the new profiles had a maximum of only four positive cues and a minimum of one positive cue (see Appendix C). In the linear environment, the regression model, the exemplar model, and the mapping model did better, on average, than the baseline model (baseline vs. EBM: $Z = -4.68$, $p < .01$). In the J-shaped environment, QuickEst and the mapping model were able to beat the baseline model (QuickEst vs. baseline: $Z = -2.05$, $p = .04$), while the exemplar model could not be distinguished from the baseline model (EBM vs. baseline: $Z = -1.37$, $p = .18$) and the regression model performed worse than the baseline model (MLR vs. baseline: $Z = -3.47$, $p < .01$).

Figure 1 illustrates the models' different successes in predicting participants' estimations. The figure shows the models' and participants' average estimations for each profile of the test phase, demonstrating that in the linear environment it is difficult to discriminate between the models, whereas in the J-shaped environment the mapping model predicted participants' estimations best. In the linear environment the regression model, the exemplar model, and the mapping model performed equally well and significantly better than QuickEst (QuickEst vs. MLR: $Z = -4.5$, $p < .01$; see also Table 5). In the J-shaped environment the mapping model was the best model in predicting the estimations (mapping model vs. EBM: $Z = -3.2$, $p < .01$) and the exemplar model was indistinguishable from QuickEst (QuickEst vs. EBM: $Z = -.03$, $p = .98$), but both the exemplar model and QuickEst outperformed the regression model (QuickEst vs. MLR, $Z = -3.59$, $p < .01$).

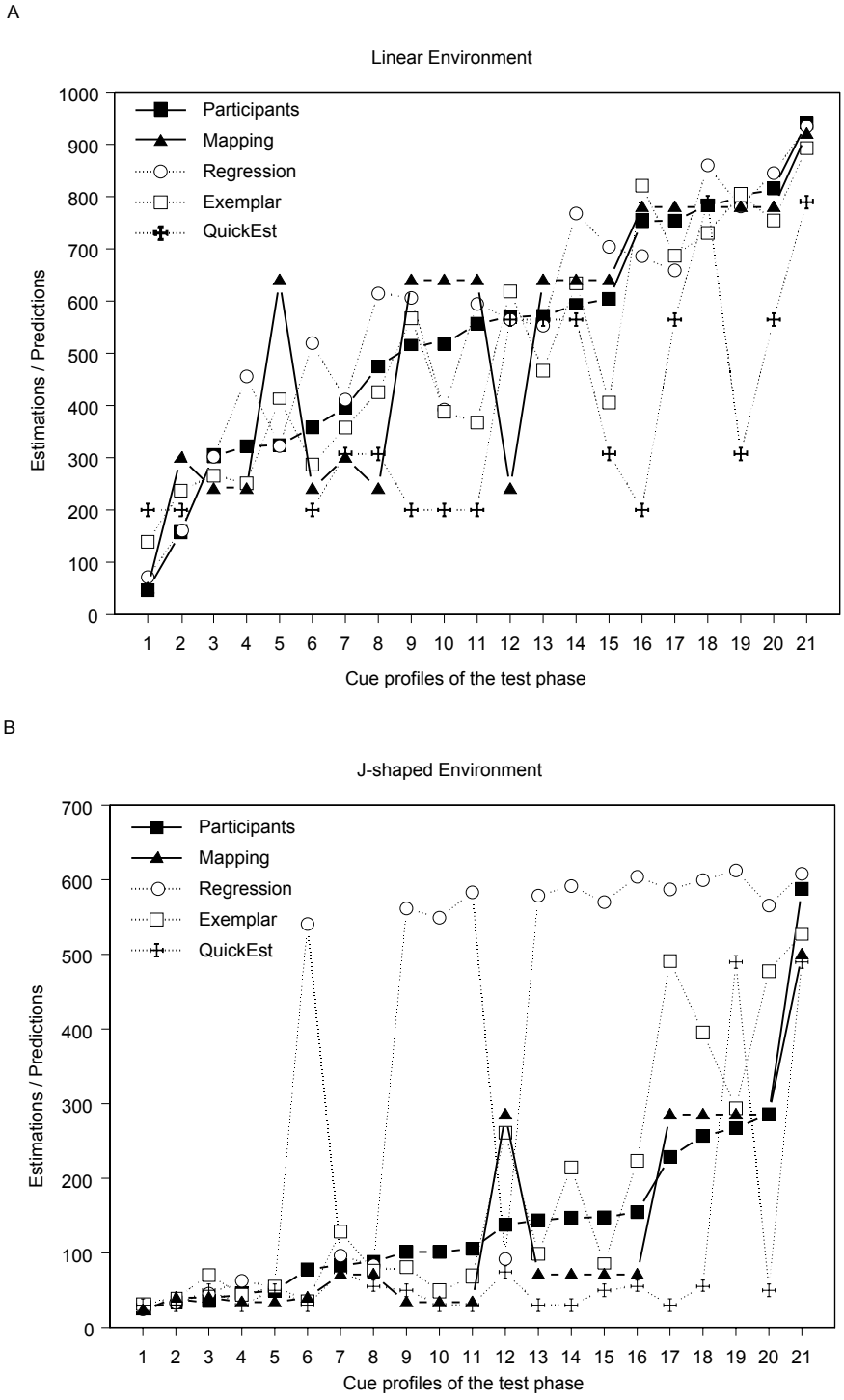

Figure 1: Models' predictions and participants' estimations in the test phase for (A) the linear environment and (B) the J-shaped environment of Study 1. The profiles in the test set are rank ordered according to the participants' average estimations. In the linear environment, profiles 1, 2, 3, 5, 7, 12, 13, and 21 were included in the test and training set. In the J-shaped environment, profiles 1, 2, 3, 4, 5, 7, 8, and 21 were included in the test set and the training set.

To consider individual differences, we examined which model, including the baseline model, was best in predicting each participant's estimations (according to the *RMSD*). In the linear environment, the mapping model was best in predicting the estimations for 12 participants (41%), the regression model was best for 11 (38%), the exemplar model for 5 (17%), and QuickEst for 1 (3%) participant. In the J-shaped environment, the mapping model was best for 16 participants (53%), QuickEst for 6 (20%), and the baseline model for 2 (7%). The regression model and the exemplar model, respectively, predicted the estimations of 3 (10%) participants best. In sum, the individual analyses led to the same conclusions as the analysis of the aggregated results: The mapping model was the best model in predicting participants' estimations. It did as well as the regression model for the linear environment, and it was the outstanding model for the J-shaped environment.

Discussion of Study 1

Study 1 showed that the mapping model was able to predict participants' estimations well in both environments, suggesting that it could be a simple alternative to standard estimation models. Although in the linear environment all models performed equally well, the exemplar model and the regression model made worse predictions compared to the mapping model in the J-shaped environment. Even though Juslin, et al. (2003, in press) showed that the exemplar model performed well in a related task, in our study, people apparently did not rely on an exemplar-based estimation process. However, this conclusion needs to be limited to the experimental situation considered, which might have been disadvantageous for an exemplar-based process. In particular, some of the cue profiles in the experiment were indistinguishable. Although this is a realistic feature in quantitative estimations in everyday life it could nevertheless have impeded an exemplar-based inference process, by making it more difficult to establish memory traces for the exemplars. Therefore in Study 2 an experimental situation was created that should favor an exemplar-based inference process and should increase the differentiability of the models in the linear environment.

Study 2

The first goal of Study 2 was to examine the reasons for the poor performance of the exemplar model in Study 1. As described above, using objects with identical cue profiles but with different criterion values could have made memorization of exemplars cognitively very

demanding. Therefore in Study 2 each cue profile appeared only once in the training set. Additionally, the objects (i.e., bugs) were given names to emphasize that the same objects had to be evaluated several times. This procedure made memorization of exemplars easier and fostered an exemplar-based inference process. It also allowed, in principle, perfect performance in the training phase, when following an exemplar-based estimation process. Thus, Study 2 provided good conditions for the exemplar model.

To test the exemplar model against the mapping model, in the test phase of the experiment all possible cue profiles that could be created with the limited number of cues were presented to the participants, so that some of the profiles had been encountered before in the training phase and some were new. The objects of the training set were presented with new names in the test phase to test for memory effects of the pure cue profiles, excluding memory effects due to memorizing exemplars by their names. To examine the consistency in estimations all profiles were presented twice, again with new names at the second appearance. This allowed us to compare the consistency in estimations for the old profiles encountered in the training phase with the consistency for the new profiles. Larger consistency for known profiles than for new profiles would indicate that memory processes played an important role in the estimations, whereas no differences between old and new profiles would speak in favor of a rule-based approach, described, for instance, by the mapping model. Finally, in Study 2 we aimed for an increased discrimination between the models' predictions.

Method

Participants. In Study 2, 50 participants took part and were randomly distributed to the two conditions, balanced for gender; 25 were women and 25 were men. The mean age was 25 years and the participants were mostly recruited from the universities in Berlin. Participants were paid according to their performance with an average payment of €17 for an individual session lasting on average 1.5 hr.

Design, procedure, and materials. The procedure of Study 2 was similar to that of Study 1, in that participants solved the same estimation task. In contrast to Study 1, the participants only had to learn 19 bugs in the training phase and had to estimate 64 (2 × 32) bugs in the test phase. They were told that in the training phase the same 19 bugs would appear 10 times each, whereas in the test phase they would have to evaluate unknown bugs. To ensure that the participants would recognize the bugs when they reappeared, each bug

received a male German name. The names were randomly assigned from a list of the most common German names. Otherwise the procedure was the same as in Study 1. Participants were paid according to the accuracy of their estimations. A similar feedback algorithm to that from Study 1 was used, with the correction terms based on the fit of the baseline model (for details see Appendix B).

In Study 2 the training set and the test set were selected in a similar way to Study 1, though with different constraints. The main objective was to improve differentiation between the mapping model, the regression model, and the exemplar model in the linear environment and the mapping model and QuickEst in the J-shaped environment. This was limited, however, by the restriction of unique profiles in the training set. Additionally, in Study 2 the correlation of the cues with the criterion was the same for the linear and the J-shaped environment (but the cue–criterion correlations differed substantially within the environments). Because in Study 1 this correlation differed between the environment conditions, this could explain why the participants differed in their accuracy of estimating the bugs' toxicity in the linear and the J-shaped environment. These changes led to slightly different training sets for the two conditions.

As in Study 1 we examined how well the models predicted the criterion values in the training phase. The exemplar model estimated the criteria perfectly in both environments, due to the unique profiles. All other models did worse with the linear environment than with the J-shaped environment. In the linear environment the regression model was the second-best model, explaining 65% of the variance of the criterion (*RMSD* = 177), whereas the worst model, QuickEst, explained only 32% of the variance (*RMSD* = 269). In the J-shaped environment, the mapping model reached the second-best accuracy for estimating the criterion values, explaining almost 90% of the variance (*RMSD* = 78), and the regression model was the worst (*RMSD* = 143, r^2 = .60). In sum, the models' accuracies differed substantially for the training phase, which can be explained by two factors. First, we created a task structure that kept the cue–criterion correlations in the linear and the J-shaped environment equal. Second, items were selected such that the differences between the models' predictions for the test phase were increased. Both factors increased the differences of the models' accuracies in the training phase.

Results

Overall, we were able to replicate the results of Study 1. The mapping model was again the best model for predicting participants' estimations when both conditions were considered jointly, and it outperformed all other models in the J-shaped environment. The exemplar model, however, did not substantially profit from the changes in the experimental structure, suggesting that exemplar-based estimation processes do not occur very frequently.

Accuracy and consistency of participants' estimations. The accuracy of the participants was measured in the same way as in Study 1 with the *RMSD* between the participants' estimations and the criterion. The participants mastered the estimation task very easily. The mean *RMSD* dropped in the linear condition from 279 in the first block to 148 in the 10th block. In the J-shaped environment the accuracy increased from an almost equally high error in the first block (*RMSD* = 215) to an *RMSD* of 51 in the 10th block. Just as in Study 1 the data from the three last blocks was merged to analyze the performance. The average *RMSD* in the linear environment was three times as high as in the J-shaped environment [$RMSD_{\text{linear}}$ = 164 vs. $RMSD_{\text{J-shaped}}$ = 58; $U = 68$, $p < .01$]. Likewise, the achievement measured by the Pearson correlation between the criterion and the estimations was on average $r = .82$ in the linear environment and $r = .96$ in the J-shaped environment ($U = 87$, $p < .01$). In sum, participants' different accuracies in the two environments reflect the environments' different difficulties.

The cue profiles of the test phase were split into two groups, one consisting of the old profiles known from the training phase and the other containing only new profiles (Table 6). To investigate participants' consistency, the correlations (and the *RMSD*) between the two estimations for the same profile presented twice in the test phase were determined. The participants were equally consistent in the two environments in their estimations for the old profiles, $r_{\text{linear}} = .90$ vs. $r_{\text{J-shaped}} = .89$; $U = 239$; $p = .16$, but the estimations for the new profiles were less consistent in the linear environment than in the J-shaped environment, $r_{\text{linear}} = .67$ vs. $r_{\text{J-shaped}} = .78$, $U = 207$; $p = .04$. The consistency for the new profiles was significantly lower than the consistency for the old profiles, $r_{\text{new}} = .72$ vs. $r_{\text{old}} = .90$; $Z = -5.03$, $p < .01$. The higher consistency in the J-shaped environment indicates that participants relied more on rule-based processes in the J-shaped environment than in the linear environment. However, the drop in consistency from the old profiles to the new profiles suggests memory effects, as the application of rules should not be influenced by the familiarity of the profile.

Table 6: Mean Consistency of the Participants in the Test Set of Study 2

	Linear				J-shaped			
	r	SD	RMSD	SD	r	SD	RMSD	SD
Old profiles	.89	0.08	129	48	.91	0.10	89	54
New profiles	.67	0.17	146	56	.78	0.17	86	42

Note. There were 25 participants in the linear environment and 25 in the J-shaped environment.

Response times. In Study 2 we measured the response times for the estimations. Response times dropped during training from a median response of 14.7 s in the first block to 7.5 s in the 10th block. There were no significant difference between the two conditions in the training phase, Mdn_{linear} = 8.4s vs. $Mdn_{\text{J-shaped}}$ = 7.1s (U = 255, p = .27), or the test phase (U = 238, p = .15). Participants responded faster at the end of the training phase (Mdn = 7.4 s) than in the test phase (Mdn = 9.9 s; Z = −4.01, p < .01), but there was no difference in response time between old and new profiles (Z =−1.3, p = .20).

Model comparison. The fit of the models was quantified in the same way as in Study 1. Again, the data of the last three blocks of the training phase were used to estimate the models' parameters and the fitted models were employed to make predictions for the test phase. Here we focus on the model performance in the generalization test, but the models' fits in the training set can be found in Appendix E. For the generalization test the items of the test phase were split into two groups: one consisting of the old cue profiles encountered in the training phase and the other of only new profiles that had not been encountered before. We first report the results on the old profiles and then come to the decisive comparison in predicting the estimations for the new profiles. In the linear environment, the regression model was the best model for the old profiles, with a significant advantage over the exemplar model (Z = −2.70, p < .01) and the mapping model (Z = −3.16, p < .01; see Table 7 for the means). In the J-shaped environment the exemplar model and the mapping model were equally good in predicting the estimations for the old profiles in the test phase (Z = −.71, p = .47) and significantly better than QuickEst or the regression model (mapping model vs. MLR: Z = −4.32, p < .01).

However, the crucial model test consists of considering how well the models are able to predict participants' estimations for new, independent profiles. As in Study 1, the baseline model was first used as a comparison standard for model performance. For the new profiles, the baseline model reached an average fit of $RMSD$ = 213 in the linear environment and of $RMSD$ = 136 in the J-shaped environment. Although the exemplar model, the regression

model, and the mapping model were better than the baseline model in the linear environment (EBM vs. baseline: $Z = -3.32$, $p < .01$), only the mapping model beat the baseline model in the J-shaped environment (mapping model vs. baseline: $Z = -4.37$, $p < .01$). This indicates that the rather naïve baseline model might not be so bad after all. Especially in the J-shaped environment, its estimations can be quite accurate, as most of the objects have similarly low criterion values. It also resonates with research on human estimation showing that people tend to rely on the mean if they must predict new objects without further information (Helson, 1964).

When comparing the models against each other the regression model, the mapping model, and the exemplar model were equally good predictors of the participants' estimations of the new objects in the linear environment (see Table 7; mapping vs. MLR: $Z = -.18$, $p = .87$; MLR vs. EBM: $Z = -1.28$, $p = .21$). In the J-shaped environment, the results become much clearer, particularly when we focus on the new objects. The mapping model was the best model; the exemplar model came in second, performing significantly worse than the mapping model ($Z = -3.27$, $p < .01$). Both models performed distinctly better than the regression model or QuickEst. In sum, the two best models (MLR and mapping) demonstrated a quite impressive fit, coming close to the variance in participants' estimations caused by inconsistencies. This error variance provides an upper limit of the fit that can be reached by any deterministic model. Surprisingly, the exemplar model could not predict participants' estimations better than in Study 1, although Study 2 provided better conditions for a memory-based estimation process.

Table 7: Models' Average Accuracies in Predicting Participants' Estimations in the Test Phase of Study 2 (Test Set)

	Linear				J-shaped			
	Mapping	Regression	QuickEst	Exemplar	Mapping	Regression	QuickEst	Exemplar
				Old				
RMSD	160	139	244	165	92	156	147	88
SD	35	36	33	35	26	9	24	31
r^2	.68	.76	.33	.68	.84	.54	.69	.85
SD	0.13	0.11	0.09	0.12	0.11	0.10	0.14	0.11
				New				
RMSD	174	172	246	184	100	216	163	148
SD	43	58	51	42	58	34	33	24
r^2	.38	.50	.25	.37	.61	.44	.29	.50
SD	0.19	0.18	0.14	0.15	0.19	0.14	0.22	0.19
				Total				
RMSD	167	154	246	174	99	186	156	118
SD	34	44	35	32	13	17	21	18
r^2	.60	.67	.27	.58	.77	.36	.44	.70
SD	0.13	0.14	0.09	0.15	0.13	0.08	0.11	0.09

Note. There were 25 participants in the linear environment and 25 in the J-shaped environment.

Qualitative analyses. The mapping model proved itself as a valid competitor with the other models. However, to enhance this conclusion drawn from the quantitative model comparison, it is desirable to provide additional qualitative support. The predictions of the mapping model are based on typical criterion values abstracted during the training phase. The mapping model assumes that this typical criterion value is the median criterion value of objects with the same cue sum. Thus the criterion value of some objects in the training set will coincide with the typical criterion value of the mapping model (or be very close to it, if the median is not defined but the mean of the two adjacent objects is used), while criterion values of others will be clearly different from the typical criterion value. According to the mapping model, objects with criterion values close to the typical criterion value should be estimated more accurately than objects with criterion values differing substantially from the typical value.

In the linear environment, this hypothesis is also compatible with estimations based on the regression model, but in the J-shaped environment the mapping model is the only model that predicts a difference in accuracy between the estimations for objects with typical criterion values and objects with non-typical criterion values. To test this hypothesis, the average errors made on typical objects (objects with the typical criterion value or the two objects with adjacent criterion values) were compared with the errors made on the non-typical objects (all other objects) in the last three blocks of the training set. In both environments the participants made significantly fewer errors estimating the criterion values for objects with typical criterion values than for objects with non-typical criterion values [linear: $RMSD_{\text{typical}} = 127$, $SE = 13$; $RMSD_{\text{non-typical}} = 179$, $SE = 17$; $t(24) = 22.90$, $p < .01$; J-shaped: $RMSD_{\text{typical}} = 38$, $SE = 6.7$; $RMSD_{\text{non-typical}} = 54$, $SE = 7.6$; $t(24) = 2.4$, $p = .03$). These results give further support to the mapping model.

Discussion of Study 2

Overall, the results of Study 2 replicated those of Study 1. The mapping model was again best in predicting quantitative estimations, if both environments are considered jointly. In the J-shaped environment, it clearly outperformed the other models. It reached a fit very close to the error variance in the data and was the best model for a distinct majority of participants. In the linear environment, though, it was still not possible to decide unambiguously which model predicted the data of the participants best—the regression model, the exemplar model, or the mapping model. The differentiation between the models was complicated by the high variance in participants' estimations in the linear environment. In the training phase as well as in the test phase, the estimations showed a high degree of inconsistency. However, the inability of the participants to learn to estimate the criterion values in the linear environment accurately is interesting in itself, as it reflects the poorer ability of the regression model and the mapping model to predict the criterion in the linear environment. Only the exemplar model predicted no differences in learning between the two environments. Because the exemplar model remembers individual cue profiles, its performance is independent of the criterion distribution.

The exemplar model predicted participants' estimations quite well for the old profiles in the test phase, but this was not true for the new profiles. The good fit for the old profiles suggests that participants relied on retrieved exemplars when a cue profile of an object was recognized from the training phase. Unfortunately it does not explain how the estimations for

the unknown profiles were made. Here the exemplar model seemed to offer a good description of the estimation process for only a minority of the participants.

Similarly, Juslin et al. (in press) showed in various experiments that the exemplar model described participants' behavior quite well in a "nonlinear task," while a regression model was better suited to predict participants' estimations in a "linear task." Similar to our task, the criterion distribution was linear in the linear task and J-shaped in the nonlinear task. However, Juslin et al. (in press) conceptualized the difference in the environmental structure not in terms of the distribution of the criterion but by the underlying cue–criterion relationship. The cue–criterion relationship specifies how the criterion is determined as a function of the cue values.

The form of the distribution and the cue–criterion relationship are related in the sense that if representative samples are taken, a linear cue–criterion relationship will result in a roughly linear distribution, and an exponential cue–criterion relationship in a J-shaped distribution. However a linear distribution does not have to stem from a linear function and there are many nonlinear functions that would not result in a J-shaped distribution. So far we have not specified the relationship between cues and the criterion in our tasks explicitly but have used a random procedure to generate the criterion distribution. To rule out that this impedes the predictive success of the exemplar model we conducted a third study, in which we chose an approach similar to Juslin et al.'s (in press) to create the objects' criterion values.

Study 3

In Studies 1 and 2 the item sets of the experiments were created by using randomly drawn samples from the simulation study that allowed discrimination between the models. Here the criterion value could only be predicted to some extent by a linear or nonlinear function of the cues. Therefore, to further generalize the empirical support for the mapping model, in Study 3 the criterion values were either a linear or a multiplicative function of the cue values (see Juslin et al., in press). Given the results of Studies 1 and 2 we only tested the mapping model against the strongest competing models, which are the standard regression model and the simplified exemplar model with one parameter.

Method

Participants. Forty students from Berlin universities participated in the study, 25 males and 15 females. The mean age was 24 years. The study lasted for approximately 1.5 hr and participants earned on average €17.

Design, procedure, and materials. Study 3 was constructed in the same way as Studies 1 and 2. To generalize the task to further contexts the estimation task was changed to a medical task. Participants had to estimate the probability that a patient would be cured of a fictitious disease. Participants were told that patients could receive different types of medication and that the information on which drugs a patient took would help them to estimate the criterion, that is, the probability that the patient would be cured within a year, ranging from 1 to 100%. The cues were five different drugs (labeled U, V, W, X, Y), which a patient could either receive or not receive. Participants were told that each drug on its own had a positive effect, but that there could be interaction effects between the drugs. In the linear environment the criterion (C_L) was a linear additive function of the cues (c_i):

$$C_L = 5 + 33c_1 + 22c_2 + 20c_3 + 15\,c_4 + 5c_5$$

In the J-shaped environment the criterion (C_J) was a multiplicative function of the cues:

$$C_J = 1.85 \cdot e^{C_L/25} - 1$$

For a large number of new cases in the generalization test we used a training set of only 16 profiles.

We created 20 different training–test sets that were used for both experimental conditions with 20 participants each. Again we aimed for an experimental item set with large discrimination between the models' predictions. Therefore we first created 1,000 training sets consisting of 16 randomly selected cue profiles and by using the two functions we determined the criterion values. The respective generalization sets consisted of the 16 profiles that did not appear in the corresponding training set. Next we excluded all sets in which cues correlated negatively with the criteria. Then we rank ordered the training sets according to how well they discriminated between each possible pair of models in the generalization set and chose the 20 environments that allowed maximum discrimination between all models. The experimental procedure was the same as that used in Studies 1 and 2. During a training phase consisting of 160 trials, participants learned to estimate the criterion value connected with each profile in the training set. After each trial participants received feedback on the correct criterion values and their performance. The order of

appearance was randomized as well as the assignment of the cues to the five different drugs and the order in which the drugs appeared on the screen. In the test phase each participant estimated all possible profiles two times without feedback. Participants were paid according to a feedback algorithm that was determined in the same way as in Studies 1 and 2 (for details see Appendix B).

Results

Overall, Study 3 replicated the results of Studies 1 and 2. The mapping model was clearly the best model in the J-shaped environment. However, in the linear environment the regression model outperformed the other models. Before we come to the model comparisons we report the participants' accuracy.

Accuracy of the participants. *The participants learned to estimate the criterion quite well in both conditions. In the linear environment the RMSD dropped from 18 in the first block to 7 in the 10th block and in the J-shaped environment from 29 to 7, with a rather stable accuracy in the last three blocks of the training phase. Participants' accuracy at the end of the training phase did not differ significantly between the two environments (RSMDLinear = 7 vs. RMSD J-shaped = 8; U = 176, p = .53).*

Did the participants capture the underlying function generating the criterion values? This can be seen in how well participants could predict the criterion values of the new cue profiles in the generalization set. In both environments participants were worse at estimating criterion values of patients with new drug combinations than with previously encountered combinations ($RMSD_{old}$ = 7 vs. $RMSD_{new}$ = 14, $Z = -5.3$, $p < .01$). However, they were significantly better in the linear environment than in the J-shaped environment ($RMSD_{J\text{-}shaped}$ = 16 vs. $RMSD_{Linear}$ = 12; $U = 123$, $p = .04$). This suggests that the participants in the linear environment captured the function generating the criterion values to some extent.

Model comparison. As in the preceding studies, the models were fitted on the last three blocks of the training phase for each. For the crucial model comparison test we focused on the generalization test of the test phase. In particular we compared the accuracies of the models in predicting participants' estimates of the criterion values for the new cue profiles, that is, combinations of drugs they had not seen during the training phase. Here the results were clear-cut. In the linear environment the regression model predicted participants' estimations significantly better than all other models, with the mapping model coming in second ($RMSD_{MLR}$ = 9 vs. $RMSD_{mapping\ model}$ = 14, $Z = -3.1$, $p < .01$). In the J-shaped

environment the mapping model clearly outperformed all other models ($RMSD_{\text{mapping model}}$ = 10 vs. $RMSD_{\text{MLR}}$ = 17, Z = −3.3, $p < .01$). The exemplar model and the regression model performed equally poorly. Figure 2 illustrates the accuracies of the different models in predicting participants' estimations in the generalization test.

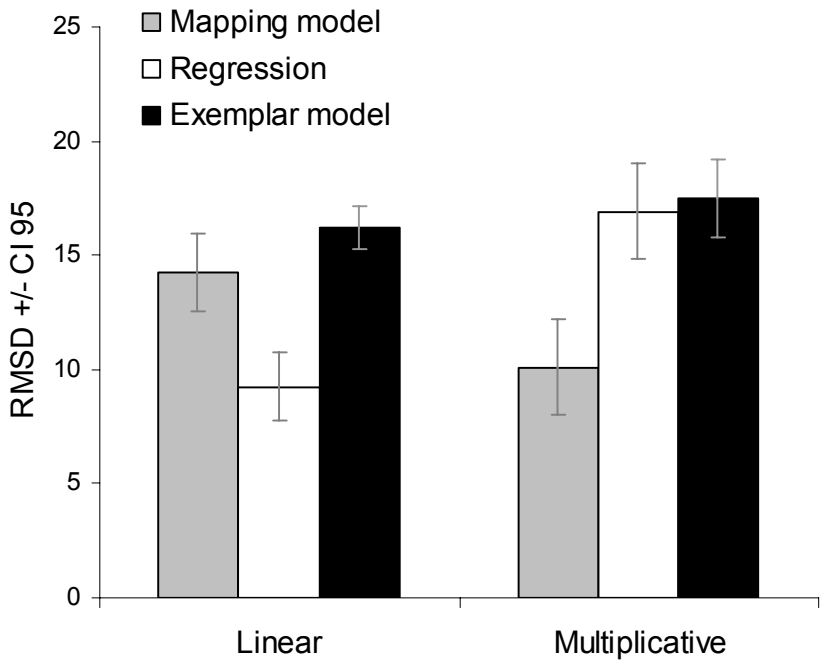

Figure 2: Models' predictive accuracies for the new profiles of the test phase of Study 3. The average root mean square deviation (RMSD) between the models' predictions and the participants' estimations for the linear and the multiplicative condition is depicted. The error bars represent the 95% confidence intervals.

Discussion of Study 3

We conducted Study 3 to test if our results from Studies 1 and 2 would also hold if the criterion distributions were generated by a linear and a nonlinear function of the cue values. In the J-shaped environment this was clearly the case. In the linear environment, however, linear regression outperformed the mapping model. As the linear criterion was generated by a linear additive function, the regression model was now equivalent to the function generating the criterion values and could estimate the criterion faultlessly. Thus if participants were able to detect the underlying structure in the data, then the regression

model would capture their estimations. We will discuss this issue further in the General Discussion.

In the J-shaped environment, we did not find a shift to an exemplar-based estimation process as advocated by Juslin et al. (in press); instead, the mapping model still described participants' behavior best. This corroborates that the mapping model is the best model for J-shaped distributions regardless of whether the underlying function has been specified.

Study 4

Study 4 represents a reanalysis of Experiment 1 of Juslin et al. (in press). In this study the authors found empirical support for a rule-based estimation process in an environment with a linear distribution of the criterion, whereas support for an exemplar-based estimation process was reported for an environment with a J-shaped criterion distribution. To test whether the mapping model, which Juslin et al. did not examine, offers an alternative account of the estimation processes, we reanalyzed the experimental data.

Juslin et al.'s experiment differed in important aspects from our studies. First, in the training phase of the experiment the participants were confronted with only 11 different profiles, a small number, that were described by only four cues, as opposed to 20, 19, and 16 different profiles with five cues each in Studies 1, 2, and 3, respectively. Second, although participants had to process less information in the training phase compared to our studies, much more training was provided by repeating each profile 20 times, as opposed to 10 repetitions in our studies. This procedure should have made it easier to memorize each profile, thus fostering an exemplar-based estimation process. Moreover, and maybe most importantly, in the experiment by Juslin et al. the participants had to learn the direction of the cues during the training phase, while in our studies the direction of the cues was told to the participants. Additionally, the cue–criterion correlations of some cues were rather small and fluctuated during training, increasing the difficulty of learning the correct direction of the cues.[6] We think this last difference is disadvantageous for a rule-based estimation process, as described by the mapping model, for which the cue–criterion correlations are essential. In sum, we think the experimental procedure is beneficial for an exemplar-based estimation process and it would be surprising if the mapping model still predicted people's behavior well.

Method

Design and procedure. The experiment had two conditions in which participants had to estimate the toxicity of bugs based on four binary cues, similar to Studies 1 and 2. Eighty participants took part in the experiment, forty in each condition. In the first condition, the linear condition, the criterion was a linear function of the cues; in the second, the multiplicative condition, the criterion was a multiplicative function of the cues. Similar to our tasks the criterion values followed either a linear or a J-shaped distribution. In an initial training phase with 220 trials participants learned to estimate the criterion values on a subset of 11 of the 16 possible bugs. In a subsequent test phase they then estimated the toxicity of all 16 bugs, that is, including the 5 bugs that they had not encountered before.

Model fitting. Following the same procedure used by Juslin et al. (in press) we fitted the models on the second half of the training data. Juslin et al. used the standard exemplar model with a free parameter for each cue, so we included this version together with the simplified exemplar model that we have reported so far. Thus, we will report results for two exemplar models, one complex exemplar model with four free parameters and one simple exemplar model with one free parameter. As in our preceding studies we analyzed the data on the individual level. We estimated the exemplar models' parameters on the second half of the training set starting with a memory base consisting of the correct cue and criterion values of the first half of the training set.[7] Then we successively added the exemplars of the second half of the training set to the memory base in the order in which they were encountered. This way the memory base always represented all objects the respective participant had seen so far (we think this method is most appropriate, because due to random error the same cue profiles had varying criterion values). The regression model was fitted directly to the participants' estimations from the second half of the training set. Consistent with our previous studies but in contrast to Juslin et al., we used an unconstrained linear regression.[8] For the mapping model we determined the directions of the cue–criterion relationship on the basis of the correlation of the cue with the criterion in the second half of the training set and then calculated the typical criterion values for each cue sum category based on the criterion values. With the estimated parameters from the training phase, each model predicted estimations for the test phase.

Results

Overall, we replicated the results of Juslin et al. (in press), but our results were not quite as clear-cut. The regression model performed best in the linear condition and the exemplar model with one parameter was the best model in the multiplicative condition. However, the simplified exemplar model was not significantly better than the regression model and the mapping model performed as well as the standard version of the exemplar model.

Model comparison. Surprisingly, and in contrast to the results of our Studies 1–3, all models performed worse in the training phase than in the test phase. In the test phase, the regression model performed best in the linear condition. It was significantly better than the mapping model and the simplified exemplar model. However, the comparison between the regression model and the standard exemplar model with four parameters only approached significance, $RMSD_{MLR} = 1.4$ vs. $RMSD_{EBM} = 1.5$, $Z = -1.78$, $p = .08$.

In the multiplicative condition it was difficult to identify one best model. The standard exemplar model used by Juslin et al. (in press) was statistically indistinguishable from the mapping model, the regression model, and the simplified exemplar model. However, the simplified exemplar model with one parameter performed slightly better than the regression model ($RMSD_{MLR} = 1.8$ vs. $RMSD_{EBM} = 1.7$, $Z = -1.65$, $p = .10$) and was significantly better than the mapping model ($RMSD_{mapping} = 2.0$, $Z = -3.1$, $p < .01$).

Discussion of Study 4

In contrast to Studies 1–3, the mapping model performed as well as or worse than the linear regression or the simplified exemplar model. This result highlights the dependence of the models' predictive accuracy on the structure of the task and indicates boundary conditions for the mapping model.

In the linear condition participants were able to pick up the linear additive structure of the task. Thus, in line with the reasoning of Juslin et al. (in press) and the results of Study 3, the regression model was the best model in the linear condition. In the multiplicative condition, however, the simplified exemplar model described participants' behavior better than the mapping model. We assume that this difference is due to the experimental procedure employed by Juslin et al., which was different from that employed in our studies. Due to a smaller number of cue profiles and more extensive training, memorization of

exemplars was presumably enhanced, favoring an exemplar-based estimation process. In contrast, the mapping model was constrained by the small number of cues and the selection of the training examples. Due to the composition of the training set the mapping model could only establish three categories, limiting the number of possible estimates to a rather small number.

However, the presumably most important difference in the tasks lies in the correlation of the cues with the criteria. In the experiment by Juslin et al. (in press), the direction of the cues had to be detected by the participants. The mapping model assumes that knowledge about the correct cue directions can be learned from the environment, but it does not specify the learning process. Thus, we assumed that participants picked up the cues' directions from the training set. However, as some cue–criterion correlations were rather small, it could easily be that some participants got the direction of the cues wrong or ignored cues that did not seem predictive for the task. In such a situation—where the direction of the cue–criterion correlation is unclear, participants have extensive experience with the exemplars, and the criterion is a nonlinear function of the cues—a shift to an exemplar-based process seems plausible. However, if all cues reliably predict the criterion and their direction is known to the participants, the mapping model seems to be the better model.

General Discussion

To describe the cognitive process underlying quantitative estimations we proposed a new cognitive theory that we called the mapping model. In four studies we tested this model against three alternative estimation models under a variety of environment conditions. We examined how well the models predicted estimations in a linear environment with a linear additive cue–criterion relationship, as opposed to a J-shaped environment with a nonlinear cue–criterion relationship.

The Success of the Mapping Model

The mapping model is built on an existing, successful framework for quantitative estimations—the so-called metrics and mappings framework (Brown & Siegler, 1993, 1996; Brown, 2002). Implementing a computational model of this framework enabled us to test the mapping model against other cognitive computational models of estimations. Naturally the way we specified the mapping model is only one possibility and there might be other and better ways to do so. Nevertheless, we think that our model captures the general idea of the

metrics and mapping framework, and when considering the empirical evidence provided by Studies 1–3, the model appears successful in predicting people's estimations. In three out of four studies, the mapping model was clearly superior to the other models in the J-shaped environment. Even in the linear environment, where a clear advantage of linear regression might have been expected, it performed equally well and was only outperformed when the criterion was perfectly predictable by a linear regression.

Rule-based Estimation

In the J-shaped environments the regression model was clearly not the appropriate model to predict participants' estimations. In the linear environments, the results were less clear. The regression model predicted participants' estimations as well as the mapping model in the first two studies but outperformed the mapping model in Studies 3 and 4. This resonates with innumerable articles that have shown that the regression model can successfully capture linear judgments (see Hammond & Stewart, 2001; Brehmer & Brehmer, 1988).

The varying results might be explained by an adaptive response to the environment (for a theoretical account see Rieskamp, Busemeyer, & Laine, 2003; Rieskamp & Otto, 2006). Because in Studies 3 and 4 the criterion values were generated by a linear additive function of the cues, the regression model was the optimal model for predicting the criterion. Thus, in their attempts to behave adaptively, the participants might have learned to follow a linear additive estimation strategy, as captured by the regression model. This adaptive response to the environment was also enhanced by the ease with which optimal cue weights of a linear additive estimation strategy could be abstracted during training. In Study 3 optimal cue weights could be reliably estimated from any pair of objects differing on only one cue. That is, when the cue changed from a negative to a positive cue value from one object to another, the criterion value always increased by a constant amount. In Studies 1 and 2, in contrast, an estimation process in line with the mapping model was equally successful. The regression model could only approximately predict the criterion and it was more difficult to judge a cue's contribution correctly. This might have favored an approach of giving equal weights to all cues, as assumed by the mapping model. It could also help explain why the regression model and the mapping model could not be distinguished in Studies 1 and 2. In a linear environment the mapping model can be equivalent to a unit weight regression model, so that the systematic variance captured by the mapping model and

the regression model potentially overlap. In addition, research on linear regression models has often shown a flat maximum effect, where equal weights lead to the same accuracy in prediction as optimal weights (Dawes, 1979; Einhorn & Hogarth, 1975).

In sum, in a task where the criterion is a linear additive function of the cues, people appear to be able to recognize the structure underlying the data and to abstract the appropriate weights for a linear additive estimation process. Consequently participants' estimations in such a situation are best described by the regression model. However, when abstracting the optimal cue weights is complicated, because the criterion is not a linear additive function of the cues, a shift to a unit weight model such as the mapping model seems to take place.

Exemplar-based Estimations

Research by Juslin et al. (2003b, in press) suggests that in the case of a linear cue–criterion relationship, rule-based processes offer a better description of human estimation than exemplar-based models. Consistently the regression model or the mapping model was best in predicting estimations when the criterion values were linearly distributed. Consistent with Juslin et al.'s (in press; Karlsson, Juslin, & Olsson, 2004) prediction that exemplar-based processes should occur for nonlinear cue–criterion relationships, we found that the exemplar model outperformed the regression model in predicting participants' estimations in J-shaped environments. However, in three of the four studies the mapping model as opposed to the exemplar model was best in predicting participants' estimations and only in Study 4 was the exemplar model best. Thus other factors besides the criterion distribution or the functional cue–criterion relationship seem to drive the models' predictive success.

The number of exemplars as well as the number of cues on which the exemplars differ and the amount of experience needed to memorize exemplars appear important: The exemplar model requires that all or at least a majority of the objects encountered during training be stored. Therefore the more information there is that has to be stored and the less often each object is encountered, the more difficult memorizing complete exemplars should become. If memorization of exemplars is difficult, a shift to a less demanding estimation process, captured by the mapping model, should be expected. Consistently we found that the mapping model performed better in Studies 1–3.

Further, the direction and the magnitude of the cue–criterion correlations and how reliably they can be abstracted when gaining experience with an estimation situation could

influence the models' predictive success. For the exemplar model the direction and the magnitude of the cue–criterion correlation is not decisive. As long as objects can be sufficiently differentiated by their cue profiles, the exemplar model will always reach perfect performance with a given set of known objects. In contrast, the mapping model relies on knowing the correct direction of the cues. However, the mapping model does not specify how knowledge about the cues' direction is acquired, instead we made the simplifying assumption that participants learn the correct direction during training. This appears justified when the cues correlate substantially with the criterion. In Study 4, however, some of the cues did not predict the criterion very well, with cue–criterion correlations fluctuating around zero, making it difficult to detect the cues' directions. This indicates that in a situation where it is difficult to abstract the direction of the relationship of the cues with the criterion and where the quality of the cues is dubitable, the exemplar model might be more suitable for predicting estimation processes than the mapping model.

In sum, the characteristics of the estimation situation of Study 4 were beneficial for an exemplar-based estimation process and detrimental for a rule-based process. We identified two task factors that influence the predictive success of the mapping model and the exemplar model in predicting estimations. We expect that the mapping model will be able to predict estimations in situations where many predictive cues are available, prior knowledge about the cues exists, and training is short. The exemplar model will be better in situations where the quality and the direction of the cues is unclear and extensive training on objects differing only on a few predictive cues is available. These expectations require further empirical tests.

Simple Heuristics for Estimation

In Study 1 a considerable number of participants were best described by QuickEst in the J-shaped environment. This raises the question under which the conditions QuickEst might capture the process of human estimation. QuickEst does not integrate information, whereas the mapping model uses all information available. For probabilistic inference tasks it has been found that models integrating information are often good predictors of people's inferences when all information is visible and easily accessible (Bröder, 2000; Newell & Shanks, 2003). In contrast, when information search is costly, shifts to simple heuristics that do not integrate information have been reported (e.g., Payne, Bettman, & Johnson, 1993; Rieskamp & Hoffrage, in press; Rieskamp & Otto, 2006). This suggests that QuickEst was in a disadvantageous position in our experiments, in which all information was presented

simultaneously on the screen. However, another recent study has also not found any empirical support for QuickEst (Hausmann, Läge, Pohl, & Bröder, in press).

Complexity of the Models

The models we considered differed in their complexity, that is, their flexibility in predicting different behaviors. Though complex models are better in fitting data, they face the problem of over-fitting—instead of describing the systematic structure of the cognitive process underlying estimation they might fit unsystematic error. To reduce the problems of model complexity in model selection we relied on a generalization test and included simplified versions of the models. To our surprise the complex standard regression model, with a free parameter for each cue, performed better than the simplified versions of the regression model. Thus, only by using its full complexity was the regression model able to predict people's estimations.

However the standard exemplar model (Medin & Schaffer, 1978, adapted by Juslin et al., 2003b), with one free parameter for each cue, apparently over-fitted the data. Overall, the simplified exemplar model, assuming that all cues are regarded as equally important, provided a better account of people's estimations. Thus the psychological interpretations of the attention parameters of the original exemplar model representing the subjective importance of each cue should be treated very cautiously. Further the linear condition of Study 4 indicates that there might be inference situations in which it becomes necessary to assume different attention weights for the cues. This leaves open the problem of predicting a priori which of the two exemplar models will predict behavior better.

The mapping model was the simplest model we considered as it entailed no free parameters and we only tested one version of it. Without flexibility, the model is unable to capture any specific ways people respond to a particular estimation situation. However, this disadvantage can turn out to be an advantage: the lack of flexibility reduces the danger of over-fitting, thereby making predictions more robust. This is particularly important because the environments we encounter in everyday life are typically noisy. For instance, environments can rarely be expressed by a linear additive function of cues, which could favor the unit weight approach taken by the mapping model. In a similar vein, the mapping model reduces the information load by ignoring the pattern of the cue values. In environments where it is unclear which cues can help to predict the quantity of interest, this might not be a utile assumption. However, if a set of predictive cues has been identified, the

assumption appears psychologically plausible, making the mapping model a good model of human estimations.

Limitations of the Mapping Model

What are the boundary conditions of the mapping model's success in predicting quantitative estimations? For one, we showed that the mapping model can be outperformed by linear regression when the criterion is a linear additive function of the cues. This touches upon a limitation of the mapping model: It relies exclusively on the objects it has encountered so far, so that—in contrast to the regression model—it is unable to extrapolate over the range of encountered criterion values. Research on function learning has shown that with sufficient practice, people are quite adept at learning a variety of one-dimensional functions (Kalish, Lewandowsky, & Kruschke, 2004; DeLosh, Busemeyer, & McDaniel, 1997; for a review see Busemeyer, Byun, DeLosh, & McDaniel, 1997). However, if multiple cue dimensions have to be integrated into a single response, the ability to extrapolate seems to be restricted to linear functions. Juslin et al. (in press; see also Karlsson et al., 2004; Olsson et al., 2006) showed in several studies that participants did not extrapolate if the cues were nonlinearly connected to the criterion. Thus, we believe that the mapping model's inability to extrapolate could to some extent mirror human behavior.

The comparison with the exemplar model in Studies 3 and 4 highlighted other boundary conditions for the mapping model. The model assumes that the direction of the cue–criterion relationship can be learned from the environment. When this is complicated the cues' direction assumed by the mapping model might not correspond with the subjective directions perceived by a decision maker, so that the predictions of the mapping model become inaccurate. Likewise, the mapping model does not specify which cues are used for the estimation process but includes all cues. In a condition where all cues are good predictors of the criterion this is a reasonable strategy, but in situations in which a few good predictors have to be picked out of a bunch of irrelevant cues, it will not work well. Further, the mapping model relies on a representative sample of criterion values for each category. In the case where a cue sum category is only represented by an outlying criterion value of one single object, the estimation of the mapping model can be distorted. Finally, another limitation of the mapping model is its application to estimation problems with only binary cues. How can the mapping model be extended to continuous cues? One way would be to dichotomize continuous cues (e.g., by a median split). However this rather crude approach

might result in an overly strong loss of information. A second possibility would be to reduce a large number of cue sum categories to a few manageable categories, for instance, by applying range–frequency theory (Parducci, 1965).

Final Conclusion

Past research on multiple cue judgments has focused on linear regression as a tool to analyze human judgments (Brehmer, 1994; Hammond, 1996). Although linear additive models can predict the outcome of estimation processes rather well, they have been criticized for not capturing the underlying cognitive process (e.g., Gigerenzer & Todd, 1999; Hoffman, 1960; see also Doherty & Brehmer, 1997). In response to this criticism, alternative estimation models have recently been proposed and tested, including exemplar models adapted to estimation problems (Juslin et al. 2003b; Medin & Schaffer, 1978), and simple heuristics such as QuickEst (Hertwig et al., 1999). Following up on the criticism, we proposed the mapping model as a simple, new cognitive theory and showed that it can successfully predict human estimation.

Appendices

Appendix A

Simulation Procedure

The simulation study examined in a 9 × 4 (in the J-shaped environment) and 9 × 6 (in the linear environment) factor design the impact of the percentage of negative cue values and the magnitude of the cue–criterion correlation. The conditions for the simulation were created in several steps. First, nine sets of five dichotomous cues with differing percentages of negative cue values were created. All cues of a set shared the same percentage of negative cue values, varying in steps of .10, between .10 and .90 per cue. The cue values were randomly assigned to the 50 objects representing an environment. Second, for each level of percentage of negative cue values we created further sets to manipulate the cue–criterion Pearson correlation. For each set with the same percentage of negative cue values we created different sets with different cue–criterion correlations. The cue–criterion correlations were varied in steps of .10 between .0 and .30 in the J-shaped environment (providing four different levels) and between .0 and .50 in the linear environment (providing six different levels). Again, all cues of a set shared the same correlation. Because the maximal possible correlation decreases with increasing percentages of positive cue values in the J-shaped environment, the number of factor levels for the correlations was lower in the J-shaped environment. The cue–criterion correlations were modified by randomly selecting two objects with different cue values and exchanging their cue values if this changed the cue–criterion correlation in the desired direction (this step was repeated until the desired correlations were obtained). This resulted in a 9 (percentage of negative cue values) × 4 (magnitude of correlation) design in the J-shaped environment and a 9 (percentage of negative cue values) × 6 (magnitude of correlation) design in the linear environment. In every condition each model was fit to half of the data and then tested on the other half.

Appendix B

Feedback Algorithms

During the training phase participants got feedback about the accuracy of their estimations and the number of points they were earning. The points participants received for their estimations were determined by the following algorithms. Any unusual deviation exceeding 500mg/l, as might be caused by a typing mistake, was treated as a deviation of 500mg/l. For each environment a different correction term (e.g., 1,100 for the linear environment in Study 1) was used to adjust for the task difficulty. The correction term was chosen dependent on the baseline model and determined the magnitude of the deviation for which a participant would receive zero points. The magnitude of the deviation that would result in zero points is given by the root of the correction term multiplied by 100. Thus in the linear environment a participant deviating less than $332 = (1{,}100 \times 100)^{1/2}$ mg/l would earn points whereas for a deviation exceeding 332 mg/l, points would be subtracted.

The equations for the feedback algorithms are defined as

$y = -x^2/c + 100$, for $x \leq 500$ and

$y = -500^2/c + 100$, for $x > 500$,

where x is the absolute difference between a participant's estimation and the actual criterion value for a given trial, y denotes the number of points that were added or subtracted from the participant's account, and c is the correction term. The correction terms for Study 1 were c = 1,100 for the linear environment and c = 550 for the J-shaped environment. The correction terms for Study 2 were c = 888.58 for the linear environment and c = 536.26 for the J-shaped environment. The correction terms for Study 3 were c = 556 for the linear environment and c = 512 for the J-shaped environment.

Appendix C

Structure of the Test Sets in Study 1

In Study 1 the test sets in the two environments differed slightly. Each test set consisted of old objects that had appeared in the training phase and new objects that the participants had not encountered before (see Tables C1 and C2).

Table C1:
Test Set in the J-shaped Environment in Study 1

Number	Profile	Cue1	Cue 2	Cue 3	Cue 4	Cue 5	Exemplar	Regression	QuickEst	Mapping
1	Old	0	0	0	0	0	23	37	30	23
2	Old	0	0	0	1	0	33	25	30	40
3	Old	0	0	1	0	1	34	32	30	34
4	Old	0	1	0	0	0	75	61	50	40
5	Old	0	1	0	1	0	41	49	50	34
6	Old	0	1	0	1	1	130	131	70	71
7	Old	0	1	1	0	1	52	56	50	71
8	New	0	1	1	1	1	284	44	70	286
9	New	1	0	0	0	0	23	559	30	40
10	New	1	0	0	0	1	29	641	30	34
11	New	1	0	0	1	0	33	548	30	34
12	New	1	0	0	1	1	232	629	30	71
13	New	1	0	1	0	1	34	554	30	71
14	New	1	0	1	1	1	566	543	30	286
15	New	1	1	0	0	0	75	584	50	34
16	New	1	1	0	0	1	242	665	50	71
17	New	1	1	0	1	0	42	572	50	71
18	New	1	1	0	1	1	317	653	500	286
19	New	1	1	1	0	1	438	579	50	286
20	New	1	1	1	1	0	566	485	50	286
21	Old	1	1	1	1	1	567	567	500	500

Note. The profiles are ordered lexicographically according to the cues' correlation with the criterion in the training set. Profiles 1–7 and 21 also appeared in the training set. The parameters for the models were set as follows: Exemplar model with one free parameter: s = .0006; regression model: intercept = 36.92, c_1 = 522.39, c_2 = 24.16, c_3 = −86.23, c_4 = −11.83, c_5 = 81.25; for QuickEst all cues were included.

Table C2:
Test Set in the Linear Environment in Study 1

Number	Profile	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Exemplar	Regression	QuickEst	Mapping
1	Old	0	0	0	0	0	50	151	200	50
2	Old	0	0	0	1	0	220	150	200	300
3	Old	0	0	1	0	1	240	244	200	240
4	Old	0	1	0	0	0	480	379	300	300
5	Old	0	1	0	1	0	307	377	300	240
6	Old	0	1	0	1	1	665	665	700	640
7	Old	0	1	1	0	1	480	472	700	640
8	New	0	0	1	1	1	240	243	200	640
9	New	0	1	1	1	1	738	470	700	780
10	New	1	0	0	0	1	145	889	200	240
11	New	1	0	0	1	0	220	600	200	240
12	New	1	0	0	1	1	608	888	200	640
13	New	1	0	1	0	1	240	69	200	640
14	New	1	0	1	1	1	920	693	200	780
15	New	1	1	0	0	0	480	829	300	240
16	New	1	1	0	0	1	686	1117	700	640
17	New	1	1	0	1	0	307	827	300	640
18	New	1	1	0	1	1	774	1115	700	780
19	New	1	1	1	0	1	810	922	700	780
20	New	1	1	1	1	0	920	632	300	780
21	Old	1	1	1	1	1	920	920	700	920

Note. The profiles are ordered lexicographically according to the cues' correlation with the criterion in the training set. Profiles 1–7 and 21 also appeared in the training set. The parameters for the models were set as follows: Exemplar model with one free parameter: s = .0001; regression model: intercept = 151.32, c_1 = 450.09 c_2 = 227.37, c_3 = −195.09, c_4 = −1.67, c_5 = 287.98; for QuickEst all cues were included.

Appendix D

Comparison of the Standard Exemplar Model, the Simplified Exemplar Model, and the Regression Model

In Study 1 we included simplified variants of the regression model and the exemplar model in addition to the standard versions. For the exemplar model we included a version with five parameters fit to each participant individually (standard exemplar), a simplified exemplar model with only one free parameter (simplified exemplar), and an exemplar model with its five parameter values optimized by using the objective criterion value instead of participants' estimations (a priori exemplar). For the regression model we included the standard model with six free parameters fit to each participant individually (standard regression), a stepwise regression model that only included the cues that received significant weights (stepwise regression), and a regression model with the parameter values optimized by using the objective criterion value instead of participants' estimations (a priori regression).

The parameters of the simplified variants of the exemplar model and the regression model were estimated in the same way as for the standard versions. For the a priori exemplar model and the a priori regression model the parameters were optimized by using the objective criterion values of the training set. The simplified exemplar model and the standard exemplar model were fitted on the last three blocks of the training phase with the correct cue and criterion values of the training set as the memory base. The best parameters for each participant were searched for by using the quasi-Newton optimization method as implemented in MATLAB. To avoid local minima, parameters were first derived by a grid search with the results serving as the starting values for the subsequent fitting procedure. The parameters for the standard regression model and the stepwise regression model were obtained by respectively determining a multiple linear regression and a stepwise regression on the last three blocks of the training set. The stepwise regression model reduced the number of employed cues substantially; on average only 3.7 (SD = 1.06) cues were used in the linear environment and only 1.5 (SD = .77) in the J-shaped environment.

Naturally, of the different versions of the exemplar and regression models, when fitted to the data of the training phase the most complex ones did significantly better than the simplified versions. In the crucial generalization test of the test phase, however, the simplified exemplar model was clearly superior to the standard version of the exemplar

model and the a priori exemplar model (all Zs < −2.48, ps < .01). The standard version of the regression model in all cases did significantly better than the two simplified versions except in the J-shaped environment, where the a priori regression model was equally as good as the standard regression model (Z = −.59, p = .57). Here we report the *RMSD*s of all versions for the test phase (see Table D1).

In Study 2 we again tested all versions of the exemplar model and the regression model in the model comparison. But similar to in Study 1, the stepwise regression and the regression with the parameters set a priori performed worse than the full model. The simplified exemplar model also performed again significantly better than the standard exemplar model and the a priori exemplar model.

Table D1
Average Predictive Accuracy of the Models in the Test Set of Study 1

	Standard exemplar	*Simplified exemplar*	*A priori exemplar*	*Standard regression*	*Stepwise regression*	*A priori regression*
			Linear environment			
RMSD	*219*	*161*	*206*	*166*	*182*	*282*
SD	*60*	*40*	*37*	*56*	*58*	*45*
			J-shaped environment			
RMSD	*242*	*166*	*179*	*342*	*359*	*352*
SD	*89*	*70*	*73*	*124*	*123*	*72*

Note. The J-shaped environment condition had 30 participants and the parameters determined a priori for the exemplar model were s_1= .0055, s_2 = .0008, s_3 = .0088 s_4 = .0005, and s_5 = .0006; the parameters for the regression model were intercept = 36.92, c_1= 522.39, c_2 = 24.16, c_3 = −86.23, c_4 = −11.83, and c_5 = 81.25. The linear environment condition had 29 participants and the parameters determined a priori for the exemplar model were s_1= .0274, s_2 = .0002, s_3 = .0049, s_4 = .0001, and s_5 = .0001; the parameters for the regression model were intercept = 151.32, c_1= 450.09, c_2 = 227.37, c_3 = −195.09, c_4 = −1.67, and c_5 = 287.98.

Appendix E

Model accuracies for the training phase of Study 2

In Study 2 all models performed better than the baseline model in predicting participants' estimations for the training phase. Because the training phase consisted of unique profiles, we expected the exemplar models to reach a fit close to the participants' accuracy. As anticipated, the exemplar model performed very well, explaining over 74% of the variance in the linear environment and 90% in the J-shaped environment. The models' accuracies are reported in Table E1.

Table E1
Model Accuracies in the Training Set of Study 2

	Environment							
	Linear				J-shaped			
	Mapping	Regression	QuickEst	Exemplar	Mapping	Regression	QuickEst	Exemplar
RMSD	192	153	253	138	83	150	144	56
SD	30	28	18	54	18	11	19	35
r^2	.58	.71	.31	.74	.88	.58	.75	.92
SD	0.12	0.09	0.06	0.14	0.05	0.06	0.11	0.08

Authors' Note

Bettina von Helversen and Jörg Rieskamp, Max Planck Institute for Human Development, Berlin, Germany. We would like to thank Peter Juslin and Linnea Karlsson for providing us with the experiment data of their previous work for a re-analysis. We gratefully acknowledge helpful comments on previous versions of this article by Peter Frensch, Peter Juslin, and Konstantinos Katsikopoulos. We would like to thank Anita Todd for editing a draft of this manuscript. Correspondence concerning this article should be addressed to Bettina von Helversen.

Bettina von Helversen
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany
Phone: (+49 30) 82406 699
Fax: (+49 03) 82406 394
Email: vhelvers@mpib-berlin.mpg.de

Footnotes

1. We chose the median as opposed to the mean to represent the typical criterion value of a cue sum category, because it provides a more robust measure of central tendency. However, the use of the median implies that in a learning situation in which the decision maker gets familiar with the estimation problem the criterion values of all encountered objects need to be stored to compute the median. In contrast, using the mean would not require storing all criterion values—the criterion value of each new object could be used to update the mean. More specifically the mean $M_{k,n}$ of all encountered objects n falling in the cue sum category k can be determined by $M_{k,n} = M_{k,n-1} + (1/n) \cdot (x_{k,n} - M_{k,n-1})$, where $x_{k,n}$ represents the criterion value of the newly encountered objects and $M_{k,n\text{-}1}$ represents the mean of all objects encountered before. Thus, this updating rule requires less demand on memory, because the decision maker only needs to store the mean and the number of objects encountered so far. In the reported studies we do not model the learning process of how people represent cue sum categories, but it is a task for future research to test whether the use of the mean as opposed to the median might have the advantage of providing a better description of the initial learning process.

2. In the case of binary cue information the multiplicative similarity rule of the original context model is a special case of a multidimensional scaling approach to modeling similarity as used by the generalized context model (Nosofsky, 1992). Thus the exemplar model we used is comparable to Nosofsky's model in how similarity is modeled.

3. According to Albers (2001), spontaneous numbers are multiples of powers of 10 $\{a \times 10^i : a \in \{1, 1.5, 2, 3, 5, 7\}\}$, where i is a natural number. They form a psychologically sensible set of coarse numbers, which increase in their crudeness as the numbers increase in magnitude (see also Hertwig et al., 1999).

4. The training and test sets in Studies 1 and 2 were selected on the basis of the predictions of the standard exemplar model. For the sake of clarity we focus throughout this article on the simplified exemplar model with one parameter—the strongest version of the exemplar model; however, the simplified versions of the models were only included post hoc. Thus the design of Studies 1 and 2 were based on the standard version of the exemplar model.

5. We used two measures of goodness-of-fit, the *RMSD* between the estimation and the criterion and the coefficient of determination (r^2). These two measures are closely related but capture slightly different aspects of the model fit. Both are based on the sum of squares error (*SSE*); but whereas the *RMSD* averages the *SSE* across the number of estimations, the coefficient of determination puts the squared error in relation to the total variance. This relationship can be demonstrated by the following equations:

$$RMSD = \sqrt{SSE(w_{LSE})/m},$$

$$r^2 = (1 - SSE(w_{LSE}))/SST$$

where *SSE* is the sum of squares error; w_{LSE} the parameter that minimizes *SSE* (w), *SST* the sum of squares total defined by $\sum_i (y_i - y_{mean})$, and m the sample size (cf., Myung, Pitt, & Kim, 2005, p. 426).

6. The cue–criterion correlations of some cues fluctuated around zero. For example, in the first three quarters of the training phase of the linear condition the third cue was positively correlated with the criterion, but in the last quarter of 55 trials it was negatively correlated with the criterion.

7. Additionally, we fitted the exemplar model in the exact same way as reported by Juslin et al. (in press) and replicated the reported fits. We chose an iterative fitting procedure to model the growing memory base during the training phase, because in Juslin et al. the criterion values were not deterministic but changed for the identical profiles due to a random error. In Studies 1, 2, and 3 the iterative fitting procedure was unnecessary due to the deterministic criterion values.

8. Our results for the regression model differ from the results reported by Juslin et al. (in press) because we implemented an unconstrained regression model. Juslin et al. restricted the intercept to be the minimum criterion value in the training set and all cue weights had to add up to 1 (see Juslin et al. in press, Appendix, p. 49). The unconstrained regression model performed better in both conditions—in particular in the multiplicative condition our results were much better than those reported by Juslin et al.

Chapter 2:
Models of Quantitative Estimations: Rule-Based and Exemplar-Based Processes Compared

Abstract

In the area of categorization it has been argued that explicit, rule-based processes and implicit, similarity-based processes compete to control behavior. A similar division of labor has been suggested for multiple-cue judgment and estimation tasks (Juslin, Karlsson & Olsson, in press). Recently, however, Helversen and Rieskamp (in press) proposed a simple rule-based model, the mapping model, that outperformed the exemplar model in a task that was thought to promote exemplar-based processing. This raised the question of under which circumstances a shift to exemplar-based processing can be observed. In the present research we investigate the impact of task structure on two core assumptions of the mapping model: the establishment of an exemplar memory base and the abstraction of explicit knowledge about the task. Our results indicate that knowledge about cues is decisive. When knowledge about cues existed, the mapping model was the best model; however, if knowledge about the cues was difficult to abstract, participants' estimations were best described by an exemplar model.

Models of Quantitative Estimations: Rule-Based and Exemplar-Based Processes Compared

How do people estimate a continuous quantity, such as the selling price of their house or the quality of a job candidate? In many cases people base their estimations on cues that are probabilistically related to the quantity being estimated. For example, when estimating the selling price of a house people could rely on information such as the size of the house, the attractiveness of the neighborhood, or the presence of a deck. Cognitive models of estimation try to explain which cues people use and how they integrate them to estimate a continuous criterion, that is, the quantity of interest.

Previous research has been dominated by the use of linear additive models for describing people's estimations, such as multiple linear regression. Recently new estimation models have been successfully introduced as alternatives to the standard regression approach. First, Juslin, Karlsson, and Olsson (in press) have argued that people frequently do not rely on rules when making estimations, but on an exemplar-based process. According to the exemplar model people estimate the criterion of an object by retrieving the criterion values of similar exemplars from memory. Second, Helversen and Rieskamp (in press) have argued that people follow a rule-based process, which differs considerably from the process assumed by linear additive approaches. Introducing the mapping model, Helversen and Rieskamp proposed that people estimate the criterion value of an object by first categorizing the object by the number of positive cue values and then using a typical criterion value of past objects with the same number of positive cues as an estimate. Although the exemplar model and the mapping model argue for conceptually different estimation processes, both models have been proposed for estimation tasks in which the standard regression approach did not provide a good account of people's estimations. The goal of the present article is to test these two models rigorously against each other to examine in more detail in which situations people follow an exemplar-based or a rule-based process for making quantitative estimations.

Models of Estimation

Consistent with the widespread assumption that human cognition comprises competing multiple systems (Ashby, Alfonso-Resese, Turken, &Waldron, 1998; Hahn & Chater, 1998; Nosofsky, Palmeri, & McKinley, 1994), models of quantitative estimations can be broadly

classified by the underlying processes they assume. In general, explicit, rule-based processes are distinguished from more implicit, similarity-based processes (Hahn & Chater, 1998; Juslin, Olsson, & Olsson, 2003; Olsson, Enkvist, & Juslin, 2006; Patalano, Smith, Jonides, & Koeppe, 2001; Nosofsky et al., 1994). The dominant approach to quantitative estimation falls clearly into the category of rule-based models. Accordingly, estimation processes are conceptualized as a process of weighting and adding information, which can be described by linear additive models such as multiple linear regression (Anderson, 1981; Brehmer, 1994; Brunswik, 1952; Hammond, 1955; Hammond & Stewart, 2001). Regression models assume that for each cue, the relation between cues and criterion is abstracted and explicitly represented as a cue weight; the judgment is then made by summing the weighted cue values. The cue weights that best describe the judgment policy are found by a regression analysis (Cooksey, 1996; Doherty & Brehmer, 1997). In this vein, linear regression has been successfully applied to analyze judgments in many areas, such as clinical diagnostics (e.g. Harries & Harries, 2001), legal and medical decision making (Ebbesen & Konecni, 1975; Wigton, 1996), and personality evaluations (e.g. Zedeck & Kafry, 1977, for a review, see Brehmer & Brehmer, 1988).

Lately, however, alternative models have been suggested to describe the estimation process (Helversen & Rieskamp, in press; Hertwig, Hoffrage, & Martignon, 1999; Juslin et al., in press). Following the idea that cognitive processes are largely a function of the characteristics of the task environment (Ashby & Maddox, 2005; Erickson & Kruschke, 1998; Gigerenzer & Todd, 1999; Juslin, Jones, Olsson, & Winman, 2003; Payne, Bettman, & Johnson, 1993; Rieskamp, 2006; Rieskamp, Busemeyer, & Laine, 2003; Rieskamp & Otto, 2006), Juslin et al. (in press) suggested a shift to exemplar-based processing in nonlinear decision tasks. In contrast to a linear estimation task, where the criterion is a linear function of the cues, a task is assumed to be nonlinear if the criterion follows from a nonlinear combination of the cues. Recently Helversen and Rieskamp (in press) proposed a new rule-based model for quantitative estimation, the mapping model, that also outperformed linear regression in a nonlinear decision task. Even more puzzling, testing the two models against each other led to inconsistent results. In the third experimental study of Helversen and Rieskamp the mapping model was clearly superior to the exemplar model, but when the mapping model was tested against the exemplar model in a reanalysis of the first experimental study of Juslin and colleagues, the results were contradictory. In our past work (Helversen & Rieskamp, in press) we presented substantial evidence to support the idea that

the mapping model is a strong competitor to the exemplar model and generated some expectations about the task characteristics leading to exemplar-based or rule-based estimation processes.

In the present article we venture to test these expectations. We claim that the nonlinearity of the environment, although important, is not a sufficient factor to trigger exemplar-based estimation processes. Instead, we argue, the two models make specific assumption about the cognitive process underlying estimation. Following these assumptions, two cognitive components of the estimation process are essential: For an exemplar-based process, the quality of the exemplar memory is essential, whereas for a rule-based process, the abstraction of explicit task knowledge is decisive. In the following, we will introduce the two models of estimation and then discuss how the structure of the estimation task affects the essential cognitive component of each model and consequently explains the diverging results reported by Helversen and Rieskamp (in press) and Juslin et al. (in press; Olsson, et al. 2006).

Competing Theories

Both Helversen and Rieskamp (in press) and Juslin and colleagues (2003; in press) argued that linear additive approaches such as linear regression can predict participants' behavior in a linear estimation task, but not in a nonlinear task. For nonlinear environments they suggested competing theories. Helversen and Rieskamp proposed the rule-based mapping model, whereas Juslin et al. suggested a similarity-based exemplar model.

The Mapping Model

The mapping model assumes a simple rule-based estimation process. Accordingly, people estimate the criterion of an object by first categorizing the object by the number of positive cue values and then using the typical criterion value of past objects with the same number of positive cues as an estimate. For example, when estimating the price of a house, the mapping model assumes that people first count the number of positive features of the house that favor a high price (e.g. great location, a deck, a swimming pool). Then the number of positive features is used to categorize the house into a certain price class and the typical price for houses within this price class is used an estimate.

The mapping model is inspired by the framework for quantitative estimation developed by Brown and Siegler (1993). Brown and Siegler proposed that two types of

information are necessary for an estimation: knowledge about the *mappings*, that is, the ordinal relation of the objects according to the criterion of interest; and knowledge about the *metrics*, that is, the numeric properties of the objects, such as the distribution, the range, or the mean of possible estimates. The mapping model relies on binary cue information; each cue is coded as having a positive or a negative cue value and all cues are coded so that they correlate positively with the criterion.

In a first step, knowledge about the mappings is inferred from the cue values by counting the number of positive cue values and grouping objects together according to their cue sums. This implies that all cues are weighted equally. In a second step, knowledge about the metric properties is derived by abstracting a typical estimate for each category, represented by the median criterion values of the objects falling into the same category. When estimating the criterion value for a new object, first the category it falls in is determined by counting the number of positive cue values, and then the typical estimate for this category is abstracted and given as an estimate.

The Exemplar Model

In contrast, the exemplar model assumes a similarity-based process. According to the exemplar model people estimate the criterion of an object by retrieving the criterion values of similar exemplars in memory. For example, when estimating the price of a house, the exemplar model assumes that people recall the selling prices of similar houses that were sold in the vicinity and use them to estimate the selling price for the house under evaluation. Exemplar models have been successfully employed to explain human behavior in categorization (Juslin et al., 2003; Kruschke, 1992; Nosofsky & Johansen, 2000). As a result of this success they were recently extended to the area of quantitative estimation (Juslin et al., 2003, in press; Olsson et al., 2006).

Exemplar models assume that estimations rely on the similarity of an object to previously encountered objects that are stored in memory. To make an estimation, these previously encountered exemplars are retrieved and compared to the probe, that is, the object under evaluation. The more the probe resembles a retrieved exemplar, the closer the estimate for the probe will be to the exemplar's criterion value. More specifically, the estimate consists of the average criterion values of the retrieved exemplars, weighted by their similarity to the probe:

$$(1)\ \hat{y}_p = \frac{\sum_{i=1}^{I} S(p,i) \cdot x_i}{\sum_{i=1}^{I} S(p,i)}$$

where $\hat{y}_p$ is the estimated criterion value for the probe *p*; *S* is the similarity of the probe to the stored exemplars; x_i is the criterion value of the exemplar *i*; and *I* is the number of stored exemplars in memory. The similarity *S* between a stored exemplar and the probe depends on how many features the exemplar and the probe share. It is calculated using the multiplicative similarity rule of the context model (cf., Medin & Schaffer, 1978), defined as

$$(2)\ S(p,i) = \prod_{j=1}^{J} d_j$$

For each cue *j* it is determined whether the cue values of the probe *p* and the stored exemplar *i* match. If they match, *d* equals one, and if they do not match, *d* equals the attention parameter s_j, which captures the impact of a cue on the overall similarity and varies between zero and one. The closer s_j is to zero, the more important the cue. If $s_j = 1$, this implies that the cue *j* is irrelevant for the evaluation of the overall similarity. The original exemplar model assumes a separate s_j parameter for each cue *j* (Juslin et al., 2003; Medin & Schaffer, 1978). However, as the original exemplar model seems to be prone to overfitting, we additionally considered a simplified version with one single attention parameter *s* for all cues (Helversen & Rieskamp, in press). In this case, *s* is an attention parameter indicating how closely a retrieved exemplar needs to resemble the probe to be considered for the estimation. The closer *s* is to zero, the more similar an exemplar has to be to the probe to have an impact on the estimation.

Model Competition

Both the exemplar model and the mapping model provide new and successful modeling approaches to quantitative estimation. However, both models were proposed to explain estimation processes in nonlinear estimation environments. Furthermore, two previous experimental studies led to rather conflicting result regarding which model provided a better account of observed estimations. In the third experimental study reported by Helversen and Rieskamp (in press), the mapping model clearly outperformed the exemplar model in predicting participants' estimations. In contrast, the reanalysis of the first

experiment of Juslin et al. (in press) as reported in Helversen and Rieskamp revealed an advantage of the exemplar model over the mapping model in predicting estimations. Surprisingly, the studies were very similar: In both studies, participants estimated a continuous criterion based on multiple binary cues. The criterion was a multiplicative function of the cues, and the participants received outcome feedback to learn the task. What factors led to these conflicting results, and how can an illuminative test of the two models be performed?

In general, human cognition can be understood as an adaptation to different environments (Ashby & Maddox, 2005; Gigerenzer & Todd, 1999; Payne et al., 1993; Rieskamp & Otto, 2006). From this view it follows that estimation processes will differ depending on the estimation situation. We argue that the conflicting results found by Helversen and Rieskamp (in press) can be explained by characteristics of the task that affect two essential cognitive components of the models: exemplar memory and knowledge abstraction. Our goal is to test the importance of these components for estimation processes.

Exemplar memory. Exemplar models assume the retrieval of previously encountered exemplars from memory. Thus the quality of memory traces for encountered exemplars plays a key role. Exemplar models can only be successfully applied if memory traces for exemplars exist and can be accurately retrieved. Differences in the quality of exemplar memory could explain the contradictory results reported in the two studies by Helversen and Rieskamp (in press) and Juslin et al. (in press). The studies differed in some aspects that potentially affected the quality of exemplar memory. For one, in the study by Juslin and colleagues the training objects were repeated twice as often as in the study by Helversen and Rieskamp. In addition, a lower number of exemplars (i.e. 11 vs. 16) and less complex exemplars (4 vs. 5 cue dimensions) were used in the training phase of the study by Juslin et al. in comparison to the study by Helversen and Rieskamp. The more often each exemplar is repeated, the better participants should be able to establish accurate memory traces. Furthermore, the fewer training objects that exist and the fewer cue values that have to be stored, the easier it should be to accurately encode and retrieve the training exemplars. Thus both factors could enhance an exemplar-based estimation process.

In line with this argument, Smith and Minda (1998) found that in a categorization task exemplar-based processes occurred later in training, while at the beginning of training participants were better described by an additive prototype model (i.e. a rule-based process). Moreover, they found that the exemplar model performed better when it learned a small

category with few dimensions than when it tackled a big, more dimensional category (Minda & Smith, 2001), suggesting that exemplar-based processes should be more accessible, the fewer training exemplars have to be learned and the more frequently training exemplars are encountered.

Knowledge abstraction. Differences in the availability of task knowledge could also explain the diverging results reported by Helversen and Rieskamp (in press) and Juslin et al. (in press). While the exemplar model relies on accurate representation of encountered exemplars in memory, the mapping model requires the abstraction of explicit task knowledge.

In the third experimental study of Helversen and Rieskamp participants were informed about the directions of the cues, providing explicit knowledge that could be directly applied in the estimation task. However, in the study by Juslin and colleagues, no prior information about the cues was given to the participants, making it more difficult to form explicit knowledge of the cue directions. For the mapping model, knowledge about the predictability and the directions of the cues is crucial. Objects can only be grouped into meaningful categories if the valid cues are used and the directions of the cues are known. Furthermore, prior knowledge about the cue directions decreases the computational demands of the mapping model and could thus foster rule-based processing. In contrast, the exemplar model, relying on the similarity relations of the objects, does not depend on any knowledge about the cues but can be applied successfully as long as objects are sufficiently differentiable. Thus, if no prior knowledge about the cues exists, this could cause a shift in the direction of exemplar-based processing. In particular, if cue directions are difficult to learn, it might be equally demanding to abstract the cue directions than to store the exemplars in memory.

In sum, the mapping model relies on the abstraction of knowledge about the cues and should profit more than the exemplar model from explicit information about the cue directions being provided. In contrast, the exemplar model seems to be particularly suited to capturing the estimation process when it is difficult to gain explicit knowledge about the cues, but intensive training and less complex training material make it possible to establish accurate exemplar memory. We investigated the influence of these factors in two studies, manipulating the quality of the memory traces established as well as the access to knowledge about the cues by varying features of the task.

Methods of Model Selection and Qualitative Tests of Models

Model selection can be a challenging task. For one, the complexity of the models needs to be taken into account. Although more complex models are better in fitting data they run the risk of overfitting; that is, they not only capture systematic variance but also fit unsystematic variance in the data. Second, models often make very similar predictions, which makes it difficult to devise tests that reliably differentiate between the models.

We addressed the problem of model selection with a twofold approach. First, similar to Helversen and Rieskamp (in press), we used a generalization test (Busemeyer & Wang, 2000). To take model complexity into account, we first estimated the models' parameters by using the data of a training phase. The estimated parameter values were then used to predict participants' estimations for new test objects. Second, we devised a qualitative test. Qualitative tests are preferable to pure quantitative models tests (Pitt, Kim, Navarro, & Myung, 2006). They are less dependent on specific parameter values and they provide a critical test of the model assumptions, providing information about the correspondence of the pattern in the data with model predictions. Therefore, we aimed to find qualitative predictions that were specific for each model and could not be derived from the competing model.

For this purpose we focused on the assumptions the models make about which objects should be treated similarly and for which objects the estimations should differ. The mapping model groups objects according to their cue sums, ignoring which specific cue has a positive value. This implies that the model treats all objects with the same cue sum alike and makes the same estimations, whereas objects with different cue sums will be treated differently and estimations will differ. The exemplar model, on the other hand, relies on the similarity relations of the objects to the stored exemplars. Thus two objects that are maximally different should also differ in which exemplars they resemble and thus in the criterion values estimated. This opens the possibility of qualitatively differentiating between the models. For example, while the mapping model will predict the same value for two objects that share the same number of positive cue values but do not match on a single cue, the exemplar model will differ in its estimations (across a wide range of parameter values).[1] We used these assumptions of the models to design qualitative model comparison tests in addition to the quantitative model comparison tests.

Study 1

The goal of Study 1 was to investigate the influence of exemplar memory on model performance. We manipulated the ease with which exemplars could be stored in memory by varying the number of training exemplars. In a multiple-cue estimation task participants evaluated job candidates according to the policy of their company. Each job candidate was characterized by six cues, which the participants could use for their evaluation. In a training phase participants were presented with a number of candidates who had been evaluated by their supervisors. Based on this training sample they could learn how their company evaluated job candidates. In a subsequent test phase, we tested how well they generalized this knowledge to new job candidates and which model was best in predicting their evaluations. We manipulated the size of the training set. In the first condition participants encountered a large number of training exemplars (24), in the second condition a small number of training exemplars (8).

Method

Participants. In Study 1, 40 participants took part, 20 in each condition. The majority of participants were students from one of the Berlin universities, with an average age of 24 years (*SD* = 4); 30% of the participants were male. Participants were randomly assigned to one of the experimental conditions, balanced for gender. The study lasted for about 1 h 45 min and participants were paid an average of €16 for their participation. One participant in the condition with a low number of training exemplars was excluded from the analysis as he did not improve in evaluating the training candidates during the training phase.

Procedure and material. The study was conducted as a computer-based experiment. The task of the participants was to evaluate the quality of job candidates for an IT position on a scale of 1 to 100 points. The more points a job candidate received the more suited he or she would be for the position. Participants received information about the job candidates on six cues and each cue could have two possible characteristics (i.e. cue values). The six cues and their binary characteristics were knowledge of programming languages (C++ vs. Java), knowledge of foreign languages (French vs. Turkish), additional skills (SAP (a software system) vs. web design), previous work experience (software development vs. system administration), previous employment area (business vs. academia), and knowledge of operating systems (UNIX vs. Windows).

Participants were told which of the two possible characteristics of the cues matched the companies' demands; characteristics that matched the companies' preferences were marked in green, while characteristics that did not meet the companies' requirements were marked in red. During training participants learned how many points job candidates with different combinations on the six cues had been awarded in previous assessments. The criterion, that is, how many points a job candidate received, was determined as a multiplicative function of the cue values (Helversen & Rieskamp, in press; Juslin et al., in press):

$$C = 0.68 \cdot e^{(22c_1 + 20c_2 + 17c_3 + 15c_4 + 14c_5 + 12c_6)/20} \quad (12)$$

where C is the points the job candidates received and c_1 to c_6 the values on the six dimensions. A positive characteristic of a cue was coded with a cue value of 1 and a negative characteristic was coded with a cue value of zero. The assignment of the weights to the cues, which characteristic of a cue was coded as positive or negative, as well as the order of the cues on the screen was randomly determined for each participant.

The study consisted of two parts, a training phase and a test phase. During the training phase participants could learn the company's evaluation policy by judging job candidates who had previously been evaluated by their supervisors. In each trial participants saw and were asked to evaluate one job candidate. After each trial they received feedback about the number of points this candidate had received from his or her supervisor, how close their estimate had been, and how many points they earned in this trial (see below). Then the next candidate appeared. All training candidates were repeated 10 times, structured in 10 blocks; the order of appearance in each block was randomly determined.

We manipulated the number of training candidates in this study: In one condition the training set consisted of a large number (24) of different training candidates; in the other condition the training set comprised a small number (8) of training objects. After the training phase participants continued with a test phase in which they had to evaluate 30 more job candidates. The test phase was similar to the training phase, with the difference that participants did not receive immediate feedback about the accuracy of their evaluations and only learned how many points they had earned after they had finished the test phase. The 30 test candidates were evaluated twice. Eight of the candidates in the test phase had also appeared during training and 22 were new candidates participants had not encountered before.

Participants' payment was based on their performance. In each trial participants could earn up to 100 points depending on how accurately they estimated the quality of the job candidates. The more they deviated from the criterion the fewer points they earned. The exact number of points subtracted for a given deviation was calculated by a feedback algorithm, based on the squared deviation from the estimation to the criterion. This resulted in a rapidly decreasing number of points with less accurate estimations. Additionally, the feedback algorithm incorporated a correction term that determined the deviation that would result in a payoff of zero. It was calculated on the basis of a baseline model that always estimated the average criterion value. Any deviation exceeding the correction term led to the subtraction of points. To exclude the subtraction of a high number of points due to a typing error, the feedback algorithm was truncated. Any deviation larger than 50 was treated as a deviation of 50. A similar feedback algorithm had been successfully used by Helversen and Rieskamp (in press) to create a moderately exacting feedback environment (Hogarth, Gibbs, McKenzie, & Marquis, 1991). After the experiments points were exchanged into euros at a rate of €0.1 for every 150 points.

Selection of training and test sets. To test which model could explain the participants' behavior best we relied on a generalization test (see Busemeyer & Wang, 2000). That is, we compared which model was better in predicting the participants' estimations for a test set consisting of objects they had not encountered during training. However, we did not just compare the quantitative fits of the models, but additionally conducted a qualitative test of the models' assumptions (see Pitt et al., 2006). Qualitative tests have the advantage that they provide a critical test of model assumptions and can be constructed to be widely independent of model parameters. For this purpose we focused on two qualitative predictions that were derived from the models' assumptions about the estimation process. Due to these different assumptions of the models it was possible to derive different ordinal predictions, that is, predicted patterns of results that are qualitatively different.

First, according to the mapping model, the same value is estimated for any two objects with an equal number of positive cues, regardless of the similarity of the two objects. In contrast, if two objects are very dissimilar, that is, if they do not match on a single cue, the exemplar model's estimations should differ. For the experimental task with six cues, an estimation situation in which the mapping model makes identical predictions and the exemplar model makes different predictions occurs for objects with a cue sum of three. To clarify, for any cue profile with three positive and three negative cues (e.g. 111000, with

each number representing the cue value of one cue), the mapping model makes the same prediction for an object with the reversed cue profile (i.e. 000111). In contrast, the exemplar model will most likely make different estimation predictions, because these two objects are maximally dissimilar—that is, they do not share any cue values.

Second, we devised an additional experiment situation in which the exemplar model made similar predictions, and the mapping model made different predictions for the test objects. The mapping model makes different predictions for objects when they have different cue sums, for instance, objects with cue sums of 2 and 4. In contrast, for these objects, which necessarily share some cue values, the exemplar model can make very similar estimations. For the test phase we selected test objects with cue sums of 2 and 4 for which the exemplar model indeed made similar predictions.

To summarize, our qualitative test comprised two conditions in which the exemplar model and the mapping model made qualitatively different predictions. While the mapping model predicted a difference between the estimations for objects with a cue sum of 4 and a cue sum of 2 and no difference for objects with a cue sum of 3, the exemplar model made the opposite predictions. However, the strength of the qualitative predictions depends on the specific training and test objects. For instance, if all training objects had the same criterion value, it would be impossible to differentiate between the models. Accordingly, we aimed at selecting training set–test set combinations where the qualitative predictions of the two models would differ as widely as possible.

We first selected the training set–test set combination for the condition with 24 exemplars. To ensure that the training set would well represent the total set, we constrained the selection of training objects to contain objects with all possible cue sums approximately in proportion to the frequency in the whole set: Each sample had to contain one object with a cue sum of 0, two with a cue sum of 1, five with a cue sum of 2, eight with a cue sum of 3, five with a cue sum of 4, two with a cue sum of 5, and one with a cue sum of 6.

To find a training set–test set combination for which the models made qualitatively different predictions, we generated 100 different training samples. Next, we calculated model predictions for the remaining objects based on the respective training samples. For the mapping model with no free parameters the predictions could be directly determined from the training samples. In contrast, for the exemplar model we first calculated the optimal parameters to predict the criterion of the training set and then made predictions based on these parameter values. From the 100 samples we selected the training set–test set

combination for which the models differed most in their qualitative predictions. For the test set we included objects with cue sums of 2 and 4 for which the mapping model made widely different estimations but the exemplar model made similar estimations. Further, we included pairs of objects with a cue sum of 3 for which the mapping model made identical predictions but the exemplar model made different predictions. Additionally, we included some extra objects on which the models differed strongly in their predictions to enhance quantitative comparisons (see Table 8 for the test set, and see Appendix A for the training set of Study 1). Lastly, we included 8 objects in the test set that had appeared in the training set. In total, the test set consisted of 30 objects, 22 new objects selected for the qualitative tests and the additional 8 old objects.

Table 8: New test objects in the condition with a large number of training objects

Objects	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Cue 6	Criterion	Mapping	Exemplar
Test 2	0	0	0	1	1	0	3	3	7
Test 2	0	0	1	0	1	0	3	3	8
Test 2	0	0	1	1	0	0	3	3	8
Test 2	1	0	1	0	0	0	5	3	7
Test 4	1	0	0	1	1	1	16	24	8
Test 4	1	0	1	1	0	1	18	24	9
Test 4	1	1	0	0	1	1	20	24	8
Test 4	1	1	0	1	0	1	21	24	8
Test 3a	0	0	0	1	1	1	5	8	6
Test 3a	1	1	1	0	0	0	13	8	26
Test 3b	0	0	1	0	1	1	6	8	2
Test 3b	1	1	0	1	0	0	12	8	25
Test 3c	0	0	1	1	0	1	6	8	3
Test 3c	1	1	0	0	1	0	11	8	14
Test 3d	0	1	0	1	1	0	8	8	14
Test 3d	1	0	1	0	0	1	9	8	24
Test 3e	1	0	0	0	1	1	7	8	3
Test 3e	0	1	1	1	0	0	9	8	27
Test/extra	1	0	1	0	1	1	17	24	10
Test/extra	0	1	1	0	1	1	16	24	16
Test/extra	0	0	0	0	1	0	1	2	3
Test/extra	1	0	1	1	1	1	37	44	100

Note. Test 2 denotes objects with a cue sum of 2, Test 3 objects with a cue sum of 3, where pairs with the same letter indicate opposite cue profiles, and Test 4 objects with a cue sum of 4. Test/extra indicates objects that were additionally included in the test set to increase the differences in model predictions.

To select the training set–test set combination for the condition with eight exemplars we repeated the procedure described above. To make conditions more comparable, the 100 training sets with 8 training objects were randomly drawn from the condition with 24

training objects, with the restriction that the training sample contained one object with a cue sum of 0, 1, 2, 4, 5, and 6, and two objects with a cue sum of 3. Again, we obtained model predictions for the remaining objects and selected a test set that maximized the differences in qualitative predictions. Again, the test set consisted of 22 new objects that were not included in the training set and the 8 known objects from the training phase. The training and the test sets are reported in Appendix A.

Finally, we explored the prediction of the models for the models' parameter space, to determine the range of parameter values for which the models make qualitatively different predictions. For the mapping model this is a simple task because it has no free parameters, so that it makes one single prediction for a specific object. In contrast, the exemplar model's predictions depend on its values for the attention parameter. We covered the parameter space of the attention parameter *s* by using the values .001, .1, 2, .5, .7, and .9. Figure 3 illustrates how the predictions of the exemplar model change with increasing parameter values.

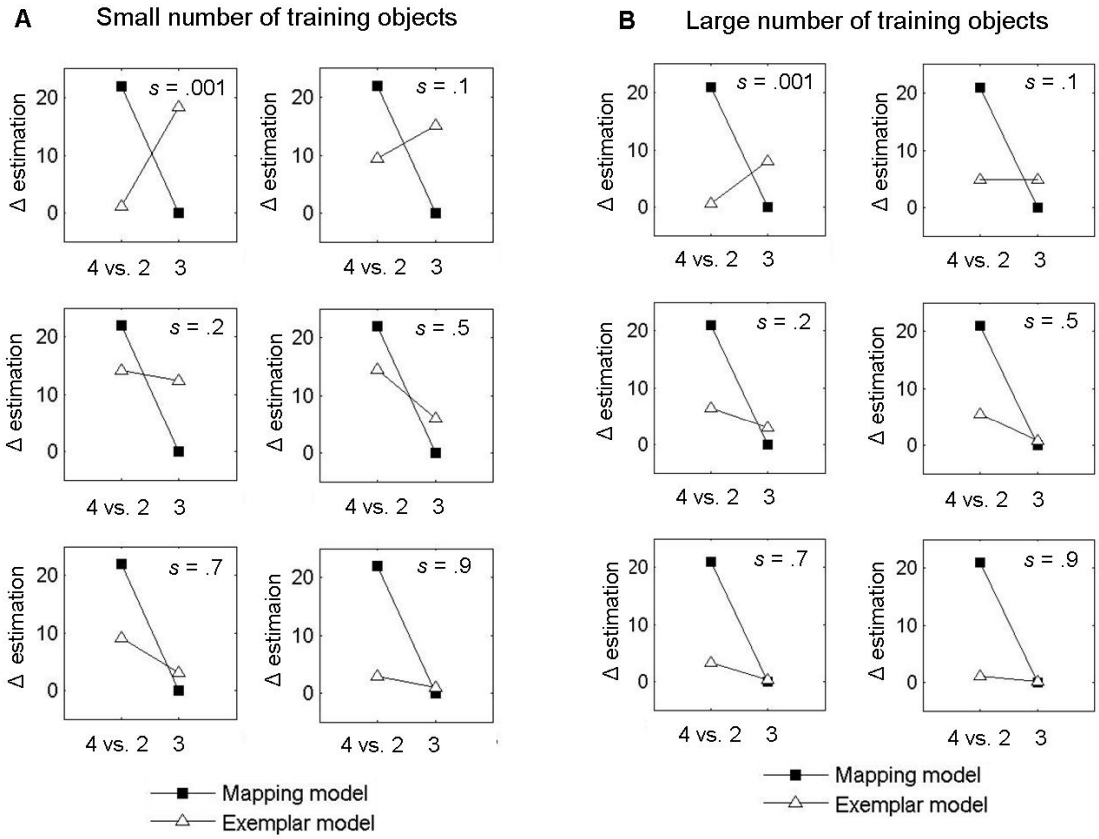

Figure 3: Qualitative model predictions. The models' predictions for the two qualitative tests, when varying the values of the exemplar model's attention parameter *s*. The "4 vs. 2" denotes the predicted average differences in estimations for the criterion values of

test objects with a cue sum of 4 and test objects with a cue sum of 2. The "3" refers to the predicted average differences in estimations for the criterion values of the pair of test objects with a cue sum of 3, with maximally different cue profiles (e.g. 111000 and 000111). (A) The predictions for the condition with a small number of training objects. (B) The predictions for the condition with a large number of training objects

With small parameter values a clear difference in the qualitative predictions of the models can be observed: In this case the mapping model predicts different estimations for objects with a cue sum of 4 versus 2, implying large differences in the estimations, which are substantially larger than the zero difference of the same estimates for pairs of objects with a cue sum of 3. In contrast, the exemplar model predicts small differences in estimations for objects with a cue sum of 4 versus 2, which are smaller than the difference of estimations for pairs of objects with a cue sum of 3. The small values for the attention parameter of the exemplar model are most plausible, because they are exactly the ones Helversen and Rieskamp (in press) found to be the best estimates for the exemplar model (i.e. the average estimated parameter values varied between .001 and .17). Thus, when assuming small attention parameter values that perform best in predicting participants' estimations, the two models make distinct ordinal predictions. Moreover, the results show that over the whole range of parameter values, the models' predictions do not overlap and that even for parameter values for which the exemplar model predicts the same ordinal data pattern as the mapping model, strong quantitative differences are to be expected.

Results

Overall, the mapping model predicted participants' estimations significantly better than the exemplar model in both conditions. Somewhat unexpectedly, the advantage of the mapping model was higher in the condition with a small number of training objects than in the condition with a large number of training objects. However, before we come to the model comparisons, we first report the participants' accuracy.

Participants' accuracy. Participants learned to evaluate the training candidates fairly well in both conditions. We measured the participants' accuracy via the root mean square deviation (*RMSD*) between the criterion values and the participants' estimations. In the condition with a large number of training objects *RMSD* dropped from 15.56, *SD* =5.62 in the first block to 3.86, *SD* = 2.07 in the 10^{th} block. Similarly, the *RMSD* in the condition with

a small number of training objects dropped from 22.97, *SD* = 6.76 in the first block to 3.04, *SD* = 4.04 in the 10th block. The participants' accuracy in the test phase did not differ between the two conditions, $RMSD_{large}$ = 5.84, *SD* = 1.87 versus $RMSD_{small}$ = 7.42, *SD* = 3.39; *U* = 137, *p* = .14. However, in both conditions, accuracy in the test phase was worse than in the training phase, $RMSD_{training}$ = 3.98, *SD* = 2.32 versus $RMSD_{test}$ = 6.61, *SD* = 2.80; *Z* = −4.16, *p* < .001. Participants were more accurate in the test phase in estimating the old objects known from the training phase than the new objects, $RMSD_{old}$ = 4.49, *SD* = 6.10 versus $RMSD_{new}$ = 6.69, *SD* = 2.21; *Z* = −3.99, *p* < .001.

Overall, participants were quite consistent in their estimations. Consistency was measured as the Pearson correlation between the first and the second presentation of the test objects. In both conditions consistency was similarly high, $r_{large}(20)$ = .95, *SD* = .06 versus r_{small} (19) = .94, *SD* = .06; *U* = 137, *p* = .14. Overall, participants were more consistent in estimating old objects than new objects, r_{old} = .98, *SD* = .05 versus r_{new} = .85, *SD* =.15; *Z* = −4.88, *p* <.001.

Model parameters. To test which model predicted participants' estimations best, we first fitted both models on the last blocks of the training phase for each participant individually. In the condition with a large number of training objects we used the last three blocks and in the condition with a small number of training objects the last four blocks to fit the models on a sufficient number of training objects. Based on the parameters estimated we made predictions for the test phase. Goodness-of-fit was determined as the *RMSD* of the model prediction from the participants' estimations. Additionally, we report the coefficient of determination r^2. The exemplar model's parameter was estimated by using participants' estimations for the last blocks of the training phase with a knowledge base consisting of the objects from the training phase with their correct criterion values. The best value for its free attention parameter was found by a grid search followed by a nonlinear least square method (as implemented in MATLAB). For the condition with a large number of training objects a mean attention parameter value of *s* = .01 (*SD* = .01) was estimated; likewise, a mean attention parameter value of *s* = .01 (*SD* = .05) was estimated for the condition with a small number of training objects. As expected, these values are rather small and correspond to the findings of Helversen and Rieskamp (in press). For the mapping model no parameters needed to be estimated. We simply calculated the typical criterion value for all objects of the training set with the same cue sum (using the correct criterion values of the objects).

In addition to the reported model comparisons, we also tested two further models to rule out that they would predict participants' behavior better than the mapping model or the simplified exemplar model we report. We included a standard exemplar model with a free parameter for every cue, as this is the exemplar model originally suggested by Juslin and colleagues (2003; see also Medin & Schaffer, 1978). We also tested a linear regression model, as it has been shown to describe participants' behavior well in other estimation tasks (Brehmer, 1994; Helversen & Rieskamp, in press; Juslin et al., in press). Both models performed worse than the mapping model and the simplified exemplar model.[2]

Quantitative model comparison. In the training set both models described participants' estimations fairly well (for means see Table 9). We used the nonparametric Wilcoxon test to analyze which model explained participants' estimations best. For the training phase the exemplar model performed better than the mapping model in both conditions, $Z_{small} = -2.20$, $p = .03$; $Z_{large} = -3.21$, $p < .01$. However, the better fit of the exemplar model during training can be explained by its higher flexibility and should not be decisive for model selection. The crucial test is how well the models predict participants' estimations in the test phase for the new objects they did not encounter during training.

Table 9: Model accuracies in Study 1

	Number of training objects			
	Large		Small	
	Mapping	Exemplar	Mapping	Exemplar
Training set				
RMSD	5.37	4.38	3.63	3.53
SD_{RMSD}	1.12	1.66	2.77	2.81
r^2	0.94	0.95	0.97	0.98
SD_{r^2}	0.03	0.03	0.05	0.24
Test set: Old				
RMSD	5.28	3.27	5.94	5.85
SD_{RMSD}	1.70	2.54	8.24	8.21
r^2	0.97	0.98	0.91	0.92
SD_{r^2}	0.02	0.02	0.22	0.22
Test set: New				
RMSD	5.87	15.45	5.74	22.63
SD_{RMSD}	2.32	2.37	3.51	1.82
r^2	0.75	0.39	0.77	0.41
SD_{r^2}	0.19	0.16	0.16	0.10
Test set: Total				
RMSD	5.80	13.39	6.63	20.00
SD_{RMSD}	1.93	2.10	4.02	2.06
r^2	0.91	0.67	0.86	0.52
SD_{r^2}	0.06	0.10	0.18	0.09

Note. $N_{\text{Total}} = 39$, with $N = 20$ in the high training condition and $N = 19$ in the low training condition. *RMSD* = root mean squared deviation

Here, the mapping model clearly outperformed the exemplar model in both conditions. In the condition with a large number of training objects it reached a *RMSD* of 5.87, *SD* = 2.32, compared to the exemplar model with a *RMSD* of 15.45, *SD* = 2.37; $Z = -3.92$, $p < .01$. Also in the condition with a small number of training objects the mapping model was clearly superior, $RMSD_{\text{mapping}} = 5.74$, $SD = 3.52$ versus $RMSD_{\text{exemplar}} = 22.63$, $SD = 1.82$; $Z = -3.82$, $p < .01$ (see also Table 9). Somewhat unexpectedly, the exemplar model performed better in the condition with a large number of training objects than in the condition with a small number of training objects ($U = 5$, $p < .01$). This appears to be contrary to the prediction that the exemplar model's performance should improve with fewer training objects. However, these results have, in fact, no implication for this prediction, because the mapping model outperformed the exemplar model in both conditions, which suggests that participants did

not rely on exemplar-based processes in either condition. Consistent with this interpretation, the mapping model performed equally well in both conditions; $U = 167$, $p = .53$.

Qualitative model comparison. Though the quantitative model comparison already indicated that the mapping model was better suited to predict participants' estimations, we additionally relied on a qualitative test. The qualitative test was designed to specifically test the models' assumptions about the cognitive process underlying estimations. To test the models' predictions, we determined for each participant and model the mean difference between the estimations for the objects with a cue sum of 2 and 4 and for the pair of objects with a cue sum of 3. As expected from the parameter space analysis illustrated in Figure 3, for both experimental conditions the models made clearly distinct qualitative predictions, as illustrated in Figure 4.

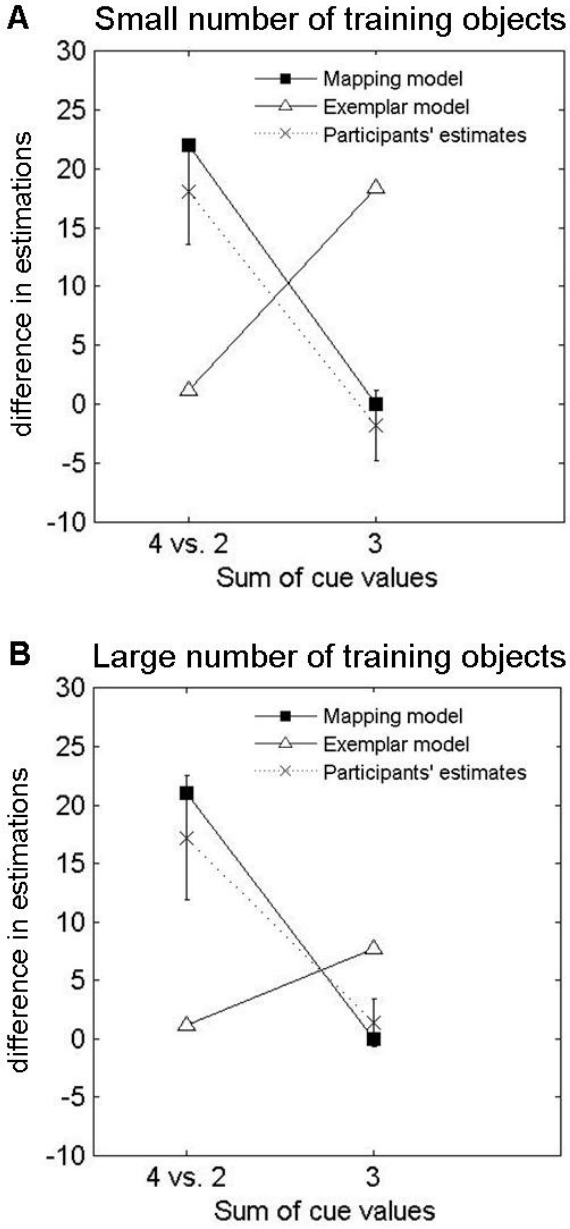

Figure 4: Qualitative test in Study 1. (A) Qualitative predictions of the models and the participants' estimations in the condition with a large number of training objects ($N = 20$). (B) Qualitative predictions of the models and the participants' estimations in the condition with a small number of training objects ($N = 19$). Sum of cue values 3 gives the average difference in estimations for the criterion values of the pair of test objects with a cue sum of 3 with maximally different cue profiles. Sum of cue values 4 vs. 2 gives the average difference in estimations for the criterion values of test objects with a cue sum of 4 and test objects with a cue sum of 2; error bars denote ±1 *SD*.

In the condition with a small number of training objects, the exemplar model predicted a small difference of 1.2 points while the mapping model predicted a difference of 22 points for test objects with cue sums of 2 and 4. In contrast, for the pairs of objects with a cue sum of 3, the mapping model predicted no difference, while the exemplar model predicted that estimations would differ by 18.4 points. Although not quite as pronounced, the same interaction was predicted in the condition with a large number of training objects. The predictions of the mapping model were clearly supported by the data. In both conditions participants' estimations differed strongly for the objects with a cue sum of 4 and a cue sum of 2. With a mean difference of 18.1 points (SD = 4.5) in the condition with a small number of training objects and 17.2 points (SD = 5.3) with a high number of training objects, they were close to the difference predicted by the mapping model. Likewise, the participants' estimations for the objects with the same cue sum but maximally different cue profiles corresponded to the assumptions of the mapping model. The difference in estimations between pairs of objects with the same cue sum were on average M = 1.3 (SD = 2.1) for the condition with 24 training objects and M = −1.8 (SD = 3.2) for the condition with 8 training objects.

Discussion of Study 1

Study 1 supported the mapping model in an estimation task with multiple predictive cues and a nonlinear criterion. It predicted well how participants estimated values for objects they had not seen during training, obviously capturing the process underlying the estimations. In comparison, the exemplar model performed quite poorly; although it was able to accurately describe the estimations during training, it could not predict the estimations for the test phase. These results indicate that the number of training objects is not a crucial factor for model performance on its own.

However, one reason we did not find an effect of the number of training objects could be that the establishment of a stable exemplar memory requires more training, even if the number of exemplars is rather small. In our study every training object was repeated 10 times, leading to a quite accurate performance of the participants in the estimation task. Nevertheless, studies investigating exemplar-based approaches often provide more training. For instance, Minda and Smith (2001) presented training objects up to 60 times each and Juslin et al. (in press) presented each object 20 times. Furthermore, Smith and Minda (1998)

suggested that exemplar-based processes only occur later in training. Thus a higher amount of training could be necessary to detect a shift in processing.

A second possibility for why the mapping model outperformed the exemplar model in both conditions is that we provided knowledge about the cue directions. This knowledge could trigger rule-based processing in accordance with the mapping model. If cue directions are known, the processes assumed by the mapping model require only a minimum of computation. However, if cue directions first need to be learned, this leads to a higher effort that has to be invested for the knowledge abstraction the mapping model requires. In contrast, for an exemplar-based estimation it is not necessary to know the direction of a cue. The amount of computation is the same, regardless of whether the cue directions are known. Accordingly, exemplar-based processes could be favored if participants do not know the cue directions. We tested these predictions in Study 2.

Study 2

Study 1 failed to elicit a shift to exemplar-based processing. In Study 2 we addressed two possible reasons for the poor performance of the exemplar model in Study 1. For one, establishing reliable exemplar memory could require extensive training. Thus, we increased the training to 20 blocks, to ensure that stable memory traces could be established. Second, the availability of explicit knowledge about the cues could have primed rule-based processing in Study 1. Because the exemplar models' performance is largely independent of explicit task knowledge, providing no information about the cues should present conditions favorable for the exemplar model. However, a shift to exemplar-based processing might not only depend on the availability of knowledge, but also on the ease with which knowledge can be abstracted. If picking up the cue directions during training is easy, the mapping model could still prevail. In Study 1 (see Table 10) all cues correlated substantially with the criterion, which should make it fairly easy to pick up the cues' directions (Hoffman & Murphy, 2006; Klayman, 1988a). Thus, to additionally manipulate the ease with which the cue directions could be learned we also manipulated how demanding it was to detect the correct directions of the cues. For this purpose we created a training set in which only half of the cues were predictive whereas the other half were useless for estimating the criterion values. This should increase the difficulty of inferring the cues' directions for predicting the criterion (Brehmer, 1973).

Table 10: Cue–criterion correlations in Study 2

	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Cue 6
Criterion (six predictive cues)	0.37	0.60	0.63	0.60	0.47	0.43
Criterion (three predictive cues)	0.79	0.15	0.17	0.58	0.56	0.11

In addition, an estimation situation in which only a few cues are predictive is a difficult one for the mapping model. The mapping model assumes that all cues are included in the estimation process. Therefore if the participants learn that only a few cues are predictive and the others can be ignored, the prediction of the mapping model, which uses all cues, can become completely wrong. Thus, if no knowledge about the cues is available and in addition it is demanding to abstract this knowledge in the training phase, this should provide optimal conditions to observe a shift form a rule-based to an exemplar-based estimation process.

Method

Participants. In Study 2, 80 students from one of the Berlin universities participated (average age = 25 years, *SD* = 3); 33% of the participants were male. Participants were randomly assigned to one of the four experimental conditions, balanced for gender. The study lasted for about 1 h 30 min and participants were paid on average €14 for their participation.

Design, procedure, and material. In Study 2 we increased the training phase, providing twice as many learning trials in comparison to Study 1. In addition, we manipulated the prior knowledge about the directions of the cues and the ease with which the cues' directions could be learned with two between-subjects factors, providing a 2 × 2 experimental design. Similar material to Study 1 was used. Again, participants were asked to evaluate the quality of job candidates based on the six binary cues described in Study 1. However, in Study 2 only half of the participants were told which cue values were regarded as positive and which as negative. The other half needed to discover the cues' directions during the training phase. Additionally, we manipulated how easily the cues' directions could be learned. One half of the participants were provided with the identical set of training objects used in the training phase of the condition with 8 training objects in Study 1. For this set of training objects all cues correlated substantially with the criterion (in all cases $r > .35$). For the other half of participants we used a different set of training objects, so that three cues correlated highly with the criterion ($r > .5$) and three correlated poorly ($r <. 2$). The exact cue–criterion correlations are reported in Table 10. The selection of objects for the training

and test phases for the second condition was achieved in the same way as in Study 1 with the additional constraint on the cue–criterion correlations and the exclusion of extreme profiles (with all positive or all negative cue values, which had to be excluded to achieve the desired cue–criterion correlations).

Similar to Study 1, Study 2 consisted of a training phase and a test phase. In the training phase participants could learn the companies' policies for evaluating job candidates by observing how many points the training candidates had received from their supervisors. The training sets in both conditions consisted of eight training exemplars. In comparison to Study 1, we increased the duration of the training to 20 trials per candidate, structured in 20 blocks. In each block the eight training candidates were presented in a random order. Participants were paid contingent on their performance, based on the same feedback algorithm used in Study 1. However, to prevent participants from becoming discouraged by overly negative feedback in the beginning of the study, we truncated the feedback algorithm, similar to in Study 1. However, to counteract the higher difficulty in the conditions with no prior information, we decreased the maximum deviation: In Study 2 any deviation larger than 30 was treated like a deviation of 30. The training phase was followed by a test phase consisting of 30 objects with 22 new and 8 old objects that participants evaluated twice. The test objects were selected in the same way as in Study 1 to allow a qualitative test of the models. The training and test sets are reported in Appendix A (Tables A2 and A3). After the test phase, participants who had not been informed about the cue directions were asked to indicate which cue values went with higher criterion values.

Results

As in Study 1, the mapping model outperformed the exemplar model when the direction of the cues was known to the participants. However, when the cue direction had to be learned during training, which model predicted the participants' estimations best depended on the number of predictive cues, that is, cues that correlated substantially with the criterion. In the condition in which all cues were predictive, the mapping model was still the best model in predicting the estimations. Only in the condition in which the direction of the cues was unknown to the participants and only three cues were predictive did the exemplar model outperform the mapping model.

Participant performance. The participants learned to evaluate the job candidates correctly in all conditions, dropping from an average *RMSD* of 27.31, *SD* =12.61 in the first

block to 3.77, *SD* = 5.90 in the 20th block. However, training accuracy depended on the knowledge of the cue directions. Participants were more accurate in their estimations when they knew the cue directions (*RMSD* = 2.07, *SD* = 2.03) than when they did not (*RMSD* = 7.43, *SD* = 6.79; *U* = 364, $p < .01$). If the cue directions were known, participants did better if all cues were predictive (*RMSD* = 1.40, *SD* = 2.03) than if only half were predictive (*RMSD* = 2.75, *SD* = 1.82; *U* = 94, $p < .01$). However, if the cue directions were not known, participants performed equally well ($RMSD_{\text{three predictive cues}} = 6.11$, *SD* = 4.09 vs. $RMSD_{\text{six predictive cues}} = 8.74$, *SD* = 8.62; *U* = 193, $p = .86$). Overall, participants' estimation accuracy was better for the training phase than for the test phase ($RMSD_{\text{training}} = 4.75$, *SD* = 5.66 vs. $RMSD_{\text{test}} = 11.82$, *SD* = 5.79; $Z = -7.62$, $p < .01$).

To measure the consistency of participants' estimations we calculated the Pearson correlation between the two judgments of the same objects during the test phase. A similar pattern to that found for participants' accuracy emerged: Participants were more consistent when they knew the cue directions ($r = .92$, *SD* = .11) than when they learned them during training ($r = .81$, *SD* = .17; *U* = 448, $p < .01$). When the participants knew the cue directions, the number of predictive cues did not matter ($r_{\text{three predictive cues}} = .92$, *SD* = .11 vs. $r_{\text{six predictive cues}} = .92$, *SD* = .10, *U* = 193, $p = .86$). However, when the cue directions were learned during training, participants were more consistent when all cues were predictive ($r = .86$, *SD* = .15) than when only three cues were predictive ($r = .76$, *SD* = .17, *U* = 122, $p = .04$). Overall, participants were more consistent in estimating the old objects than estimating the new objects ($r_{\text{old}} = .93$, *SD* = .14 vs. $r_{\text{new}} = .79$, *SD* = .22; $Z = -5.50$, $p < .01$).

Knowledge of cue directions. To examine whether our manipulation of the ease with which the cue directions could be learned had an effect, we compared how many mistakes participants made in reporting the correct directions of the cues. As expected, participants performed better when all six cues were predictive (i.e. correlated substantially with the criterion) than when only three cues were predictive. When all cues were predictive, 7 (35%) participants indicated for at least one cue an incorrect direction; whereas when only three cues were predictive, 14 (70%) participants made at least one mistake. In particular, the participants had difficulty in correctly reporting the direction of the low-quality cues (i.e. those that correlated only slightly with the criterion), with a total of 16 mistakes in comparison to only 8 mistakes with the high-quality cues.

Quantitative model comparison. As in Study 1 we used the last four blocks of the training phase to estimate individually the exemplar models' attention parameter.

Furthermore, we used the objects' correct criterion values in the training phase to determine the median estimates for the mapping model's estimation categories. The categories were formed on the basis of all six cues.[3] In this way we determined the models' predictions for the new objects in the test phase. Model performance was measured as the *RMSD* between model predictions and participants' estimations. Additionally, we report the r^2 as a second measure of goodness-of-fit. Again, we also included a version of the exemplar model with a free parameter for every cue and a linear regression model in the comparison. As neither of the two models was the best model in any condition, we do not report the model fits here. Although our conclusions are not affected by these results, they are reported in Appendix B to provide a complete picture.

Again, the exemplar model ($RMSD_{total}$ = 4.51, SD = 5.29) outperformed the mapping model ($RMSD_{total}$ = 4.89, SD = 5.54) during training in describing participants' estimations, Z = −5.25, p <.01. This was expected, as the exemplar model is more flexible than the mapping model. Interestingly, both models fitted the participants better when participants knew the cue directions than when they did not, $U_{mapping}$ = 375, $U_{exemplar}$ = 364, both p < .01; following a similar pattern to the accuracy of the participants. Table 11 reports for all conditions the mean *RMSD*s and *SD*s.

Table 11: Model accuracies in Study 2

	Number of predictive cues							
	Six predictive cues				Three predictive cues			
	Cue directions				Cue directions			
	Known		Unknown		Known		Unknown	
	Mapping	Exemplar	Mapping	Exemplar	Mapping	Exemplar	Mapping	Exemplar
Training set								
RMSD	1.77	1.40	8.86	8.40	2.89	2.68	6.03	5.61
SD_{RMSD}	1.77	2.03	8.56	8.07	1.75	1.78	3.94	3.63
r^2	.99	.99	.88	.87	.93	.94	.74	.74
SD_{r^2}	.01	.01	.17	.20	.07	.07	.27	.26
Test set: Old								
RMSD	3.44	3.25	8.39	8.92	3.12	2.99	6.35	6.02
SD_{RMSD}	3.73	3.90	7.94	8.05	2.46	2.44	3.22	3.03
r^2	.98	.98	.89	.87	.92	.92	.72	.72
SD_{r^2}	.06	.06	.19	.20	.11	.11	.23	.23
Test set: New								
RMSD	6.34	23.50	16.34	22.22	10.36	14.78	12.24	8.71
SD_{RMSD}	4.00	2.85	7.36	4.29	4.50	3.47	2.21	1.92
r^2	.71	.37	.43	.40	.66	.24	.17	.39
SD_{r^2}	.27	.17	.25	.23	.17	.09	.16	.16
Test set: Total								
RMSD	5.98	20.29	14.88	19.90	9.11	12.80	11.09	8.16
SD_{RMSD}	3.47	2.48	7.02	4.20	3.86	3.01	1.95	1.94
r^2	.89	.51	.64	.52	.69	.36	.28	.49
SD_{r^2}	.12	.11	.23	.19	.16	.09	.16	.18

Note. $N_{\text{Total}} = 80$, with $N = 20$ in each condition. *RMSD* = root mean squared deviation

Again, the crucial model comparison test consisted of how well the two models were able to predict participants' estimations for the new independent objects of the test phase. Figure 5 shows that in the condition replicating the Study 1 condition with a small number of training objects (where the participants knew the cue directions and where all cues were predictive), with the only difference being having a larger number of training trials, the mapping model again clearly outperformed the exemplar model, $RMSD_{\text{mapping}} = 6.33$, $SD = 4.00$ versus $RMSD_{\text{exemplar}} = 23.50$, $SD = 2.85$, $Z = -3.92$, $p < .001$. Thus, by simply having more training, the participants did not switch to an exemplar-based estimation process. Similarly, when the cue directions were known but only half of the cues were predictive, the mapping model predicted the participants' estimations better than the exemplar model,

$RMSD_{mapping}$ = 10.36, SD = 4.50 versus $RMSD_{exemplar}$ = 14.78, SD = 3.47, Z = −3.92, p < .01. Furthermore, the mapping model was still the superior model when the participants had to learn the directions of the cues, and all cues were predictive, $RMSD_{mapping}$ = 16.34, SD = 7.36 versus $RMSD_{exemplar}$ = 22.22, SD = 4.29, Z = −2.80, p < .01. However, when the participants needed to abstract the directions of the cues during training and this was difficult because only three cues were predictive, the exemplar model outperformed the mapping model, $RMSD_{mapping}$ = 12.24, SD = 2.20 versus $RMSD_{exemplar}$ = 8.71, SD = 1.92, Z = −3.62, p < .01.

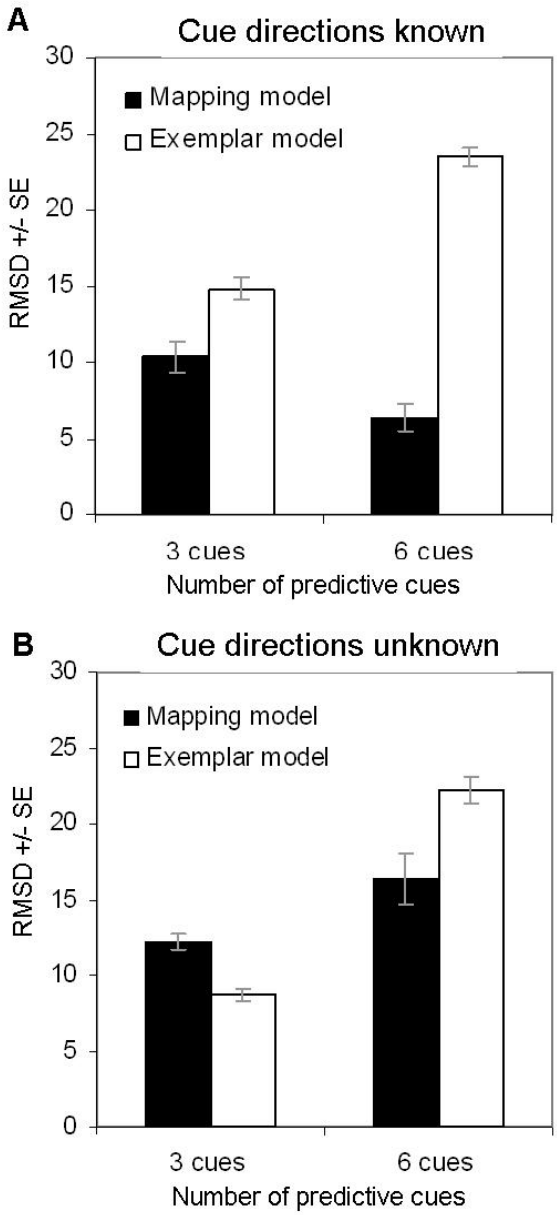

Figure 5: Models' accuracy in predicting the participants' estimations for the new objects in the test phase of Study 2. (A) Models' accuracy when the cues' directions were

known (N = 40; 20 for each condition). (B) Models' accuracy when the cues' directions were not known (N = 40; 20 in each condition).

An additional analysis of the correlation between the models' accuracy and the number of mistakes participants made when indicating the cue directions provided further evidence for a shift in processing. In the condition with unknown cue directions and six predictive cues, the mapping model performed worse the more cue directions a participant had indicated incorrectly, $r(20) = .64$, $p < .01$, suggesting that the difference in performance between the conditions with known and unknown cue directions was at least partly due to the failure of some participants to learn the cue directions. In contrast, in the condition with only three predictive cues this relation was not significant, $r(20) = .21$, $p = .38$, suggesting a shift in processing.

Qualitative model comparison. Similar to Study 1, we also tested which of the qualitatively different predictions of the two models were in line with the observed estimations. Again, we compared the predictions of the exemplar model and the mapping model by taking the difference in estimations for the pairs of objects with a cue sum of 3 and the objects with cue sums of 2 and 4. For the pairs of objects with a cue sum of 3 the mapping model made the same predictions whereas the exemplar model made different predictions. In contrast, for the objects with cue sums of 2 and 4 the mapping model made the different predictions and the exemplar model made similar predictions.

Figure 6 shows that the results of the qualitative tests clearly supported the quantitative model comparison tests. When the participants knew the cue directions, their estimations were in line with the mapping model's predictions. Similarly, when the participants did not know the cue directions, but all cues were predictive, the participants showed a similar pattern to that predicted by the mapping model. Only in the condition in which the participants did not know the cue directions and only three cues were predictive was the qualitative pattern of the estimations consistent with the exemplar model's predictions.

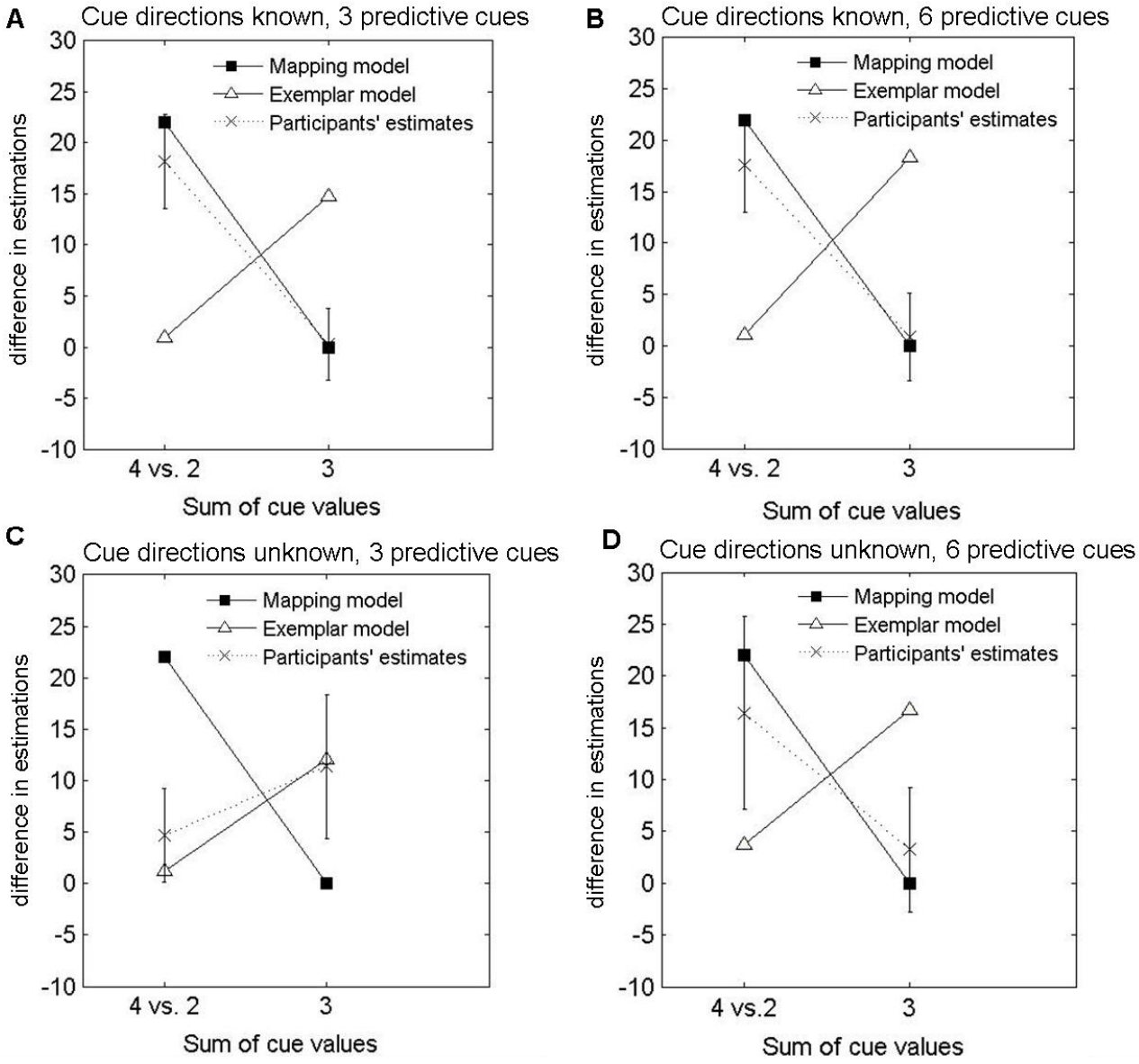

Figure 6: Qualitative test in Study 2. (A) Qualitative tests for the condition with known cue directions but only three predictive cues are shown. (B) Qualitative tests for the conditions with known cue directions and six predictive cues. (C) Qualitative tests for the condition with unknown cue directions and three predictive cues. (D) Qualitative tests for the condition with unknown cue directions and six predictive cues. Sum of cue values 3 gives the average difference in estimations for the criterion values of the pairs of test objects with a cue sum of 3 with maximally different cue profiles. Sum of cue values 4 vs. 2 gives the average difference in estimations for the criterion values of test objects with a cue sum of 4 and test objects with a cue sum of 2. Error bars denote ±1 *SD*; $N = 20$ in each panel.

Discussion of Study 2

Study 2 confirmed our prediction that a rule-based process as described by the mapping model depends on accurate knowledge abstraction, which is not necessary for the exemplar model. In line with this theoretical foundation, which model was most capable of predicting participants' estimations in Study 2 crucially depended on accurate knowledge about the cue directions. In the two conditions in which participants were told which cue values were regarded as positive evidence, the mapping model was clearly better in explaining participants' behavior. However, when the participants had to learn the cue directions during training and when this was difficult, because only three cues substantially correlated with the criterion, the exemplar model was the superior model. These results are consistent with the results reported by Helversen and Rieskamp (in press) and Juslin et al. (in press) and shed light on why the authors had found support for the mapping model in one study, while in another study the exemplar model was superior.

Although the mapping model clearly outperformed the exemplar model in the two conditions in which all cues were predictive, it should be noted that the mapping model predicted the estimations worse when the participants learned the cue directions than when the participants were informed about the directions. This result is partly attributable to some participants who failed to learn the cue directions. Furthermore, the condition where the cue direction had to be learned was apparently also quite difficult, indicated by a high variance between participants' estimations and the relatively poor performance during training. Thus, although participants managed to learn most of the cue directions, this happened at the expense of accuracy.

General Discussion

Past research has proposed that multiple distinct processing systems control human cognitive behavior. Which system wins out depends on the structure of the task (e.g. Ashby et al., 1998; Juslin et al., in press). For instance, explicit, rule-based processes are assumed to be constrained to tasks in which stimulus dimensions are separable and can be selectively attended to, while implicit, similarity-based processes catch on if the stimulus dimensions are integral (Ashby et al., 1998). Likewise, Erickson and Kruschke (1998; Ashby et al., 1998) argued that rule-based processing in categorization is restricted to easily verbalizable, uni-dimensional rules.

Following up on this line of research, our goal was to test how two recent models of quantitative estimation, the rule-based mapping model (Helversen & Rieskamp, in press) and a similarity-based exemplar model (Juslin et al., 2003; in press), are affected by different task structures. This test was based on theoretical considerations of the crucial cognitive components that are essential for the two models: the establishment of accurate knowledge abstraction for the mapping model and of accurate exemplar memory for the exemplar model. This theoretical grounding allowed us to investigate the link between cognitive processing and task characteristics. Accordingly we predicted that the mapping model would describe participants' estimations well when knowledge about the task was available or could be easily abstracted during the task. In contrast, exemplar-based processes should be triggered when knowledge abstraction is difficult but the stimulus material allows the accurate storage and retrieval of training exemplars. Our results supported our predictions. The mapping model performed best when the participants were informed about the cues' directions or could abstract them during training. However, when abstracting knowledge about the cues was difficult but exemplar memory could be used for accurate estimation, the exemplar model was best in predicting participants' estimations. In the following we will discuss the relevance of establishing accurate knowledge abstraction and exemplar memory for quantitative estimations in more detail.

Exemplar Memory: Number of Training Trials and Number of Objects

Our results showed that simply increasing the amount of training is not sufficient to trigger an exemplar-based estimation process. Even after we doubled the amount of training in Study 2 and used a small number of objects that had to be learned in the training phase the mapping model still outperformed the exemplar model in predicting participants' estimations, supporting a rule-based estimation process. Thus, when the participants had access to explicit task knowledge, no shift to an exemplar-based processing occurred even when the training intensity was increased. However, these results only hold in the cases where participants were informed about the cue directions. This suggests that the opportunity to establish stable memory traces of the exemplars does not necessarily lead to reliance on exemplar-based processes. Rather, the processes underlying estimation seem to be triggered early on: In situations in which sufficient knowledge is provided about the task structure people start with a rule-based estimation process and do not necessarily switch to an exemplar-based process even if extensive training is provided. In contrast, when only little knowledge is available about the task structure exemplar-based processes might become

more frequent and are enforced by an increased amount of training and a smaller number of training instances (Smith & Minda, 1998).

Knowledge Abstraction

Providing explicit knowledge about the cues led to a strong effect on the estimation process. The mapping model, which relies on the abstraction of explicit knowledge about the task, clearly suffered when no knowledge about the cue directions was given to the participants prior to the task. Furthermore, in both conditions the participants performed worse during training, indicating that if knowledge about the task needs to be acquired during training, learning can be impeded.

However, the exemplar model was only better in predicting participants' estimations when just a subset of the cues substantially correlated with the criterion. This suggests that lacking knowledge about the cue directions is not sufficient to trigger exemplar-based processing, but that the accuracy and difficulty with which rule-based estimation processes could be employed played an important role when a shift to exemplar-based processing occurred (see also Ashby et al., 1998; Juslin et al., in press; Olsson et al., 2006).

The condition with no prior information about the cue directions and only three predictive cues provided especially problematic circumstances for the mapping model, because it affected two of its core assumptions. First, the mapping model assumes that explicit knowledge about the cues is abstracted. This was difficult to achieve, as no information about the cues was available and the cue directions were difficult to pick up. Second, the mapping model assumes that all cues are equally important. However, in this task, in fact, only three cues were substantially correlated with the criterion. Thus, if participants learned to ignore the less valid cues (Castellan, 1973; Klaymann, 1988b), the mapping model should not be able to predict participants' estimations accurately.

This raises the question of why the mapping model performed well when only three cues were predictive and information about the cue directions was available. The good performance of the mapping model in this condition implies that participants regarded all cues as equally important for making the estimations. Thus, following a rule-based process as described by the mapping model in this condition implies that the participants did not accurately learn the task structure. Instead, to improve their estimation accuracy it would have been advantageous to use only the predictive cues for an estimation. Apparently, providing the participants with explicit knowledge about the direction of the cues led to the

inference that all cues were relevant to predict the criterion and thereby triggered a rule-based process. This finding is consistent with a "rule bias" as documented by Ashby and colleagues (1998; see also Olsson et al., 2006). This rule bias implies an initial preference for rule-based over more implicit processing, such as exemplar-based processes. In sum, our results indicate that also in quantitative estimation problems people mainly follow an exemplar-based process when a rule-based process does not provide an accurate solution to the estimation task. The theoretical considerations of the crucial cognitive components that are essential for the two models, namely, the establishment of accurate knowledge abstraction for the mapping model and of accurate exemplar memory for the exemplar model allows the prediction of under which condition a shift to exemplar-based processing can be expected.

Conclusion

Previous research has described estimation processes almost exclusively with multiple linear regression models. Recently new cognitively motivated models, such as the exemplar model by Juslin et al. (in press) and the mapping model by Helversen and Rieskamp (in press; see also Brown & Siegler, 1993) have been proposed to model estimation processes. Interestingly, these models represent two different views on estimation processes. While the exemplar model proposes an implicit, similarity-based process, the mapping model assumes a rule-based process (Ashby et al., 1998; Hahn & Chater, 1998; Juslin et al., 2003, in press). Consistent with previous research on the interplay of rule-based and similarity-based systems in categorization problems, we found evidence for an initial preference for rule-based processes in quantitative estimation tasks. Furthermore, the experimental studies reported in the present article successfully illustrate the link between the cognitive processes assumed by the models and the structure of the environments. We showed that the models' assumptions about the estimation process were directly affected by different structures of the estimation task, which consequently determined which estimation process prevailed. This highlights not only the impact of task characteristics on information processing, but also the importance of explicit assumptions about the cognitive process for computation modeling approaches.

Appendices

Appendix A

Training and test sets for Studies 1 and 2

The following tables describe the sets of items that were used in Study 1 and Study 2. Table A1 describes the set of items for the training phase of Study 1. Table A2 describes the set of items for the training and test phases of the condition with a low number of training objects in Study 1 and for the condition with six predictive cues in Study 2. Table A3 describes the set of items for the training and test phases of the conditions with three predictive cues in Study 2.

Table A1: Sets of objects for the training phases of Study 1

Training condition	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Cue 6	Criterion
A & B	0	0	0	0	0	0	1
A	0	1	0	0	0	0	2
A & B	1	0	0	0	0	0	2
A & B	0	0	0	0	1	1	2
A	0	0	0	1	0	1	3
A	0	1	0	0	0	1	3
A	0	1	0	0	1	0	4
A	1	0	0	0	1	0	4
A	0	0	1	1	1	0	7
A	0	1	0	0	1	1	7
A & B	0	1	0	1	0	1	7
A	0	1	1	0	1	0	9
A	1	0	0	1	0	1	8
A & B	1	0	1	0	1	0	10
A	1	0	1	1	0	0	10
A	1	1	0	0	0	1	10
A	0	1	0	1	1	1	14
A & B	1	1	0	1	1	0	24
A	1	1	1	0	0	1	24
A	1	1	1	0	1	0	26
A	1	1	1	1	0	0	27
A & B	0	1	1	1	1	1	33
A	1	1	1	1	1	0	55
A & B	1	1	1	1	1	1	100

Note. A & B = objects that were used for the training condition (A) with a large number of training objects and for the training condition (B) with a small number of training objects. A = objects that were additionally used in the training condition (A) with a large number of training objects.

Table A2: Sets of objects for the training and test phases of Study 1 for the condition with a small number of training objects and of Study 2 for the condition with six predictive cues

Objects	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Cue 6	Criterion	Mapping	Exemplar
Test/training	0	0	0	0	0	0	1	1	1
Test/training	1	0	0	0	0	0	2	2	2
Test/training	0	0	0	0	1	1	2	2	2
Test/training	0	1	0	1	0	1	7	8	7
Test/training	1	0	1	0	1	0	10	8	10
Test/training	1	1	0	1	1	0	24	24	24
Test/training	0	1	1	1	1	1	33	33	33
Test/training	1	1	1	1	1	1	100	100	100
Test 2	0	0	0	1	0	1	3	2	7
Test 2	0	0	0	1	1	0	3	2	9
Test 2	0	0	1	0	1	0	3	2	10
Test 2	0	1	0	0	0	1	3	2	7
Test 2	0	1	0	0	1	0	4	2	9
Test 2	0	1	0	1	0	0	4	2	7
Test 3a	0	0	1	0	1	1	6	8	2
Test 3a	1	1	0	1	0	0	12	8	24
Test 3b	1	0	1	0	0	1	9	8	6
Test 3b	0	1	0	1	1	0	8	8	24
Test 3c	1	0	0	0	1	1	7	8	2
Test 3c	0	1	1	1	0	0	9	8	20
Test 3d	1	0	0	1	0	1	8	8	5
Test 3d	0	1	1	0	1	0	9	8	21
Test 3e	1	1	0	0	0	1	10	8	5
Test 3e	0	0	1	1	1	0	7	8	21
Test 4	1	0	1	0	1	1	17	24	10
Test 4	1	0	1	1	1	0	20	24	10
Test 4	1	1	0	1	0	1	21	24	7
Test 4	1	1	1	0	1	0	26	24	10
Test/extra	1	0	1	1	1	1	37	33	100
Test/extra	1	1	1	1	0	1	50	33	100

Note. Test/training indicates the eight objects that constituted the training set in the condition with a small number of training objects in Study 1 and the two conditions with six predictive cues in Study 2. These eight objects also appeared in the respective test sets. Test 2 denotes objects with a cue sum of 2, Test 3 objects with a cue sum of 3, where pairs with the same letter indicate opposite cue profiles, and Test 4 objects with a cue sum of 4. Test/extra indicates objects that were additionally included in the test set to increase the differences in model predictions.

Table A3: Sets of objects for the training and test phases of Study 2 for the condition with three predictive cues

Objects	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Cue 6	Criterion	Exemplar	Mapping
Test/training	0	0	1	0	0	0	2	2	2
Test/training	0	0	0	1	0	1	3	3	3
Test/training	0	1	1	0	0	0	4	4	3
Test/training	0	1	1	0	1	0	9	9	8
Test/training	1	0	0	1	0	1	8	8	8
Test/training	1	1	0	1	1	0	24	24	25
Test/training	1	1	1	1	0	0	27	27	25
Test/training	1	0	1	1	1	1	37	37	37
Test 2	0	1	0	0	1	0	4	9	3
Test 2	1	0	0	0	0	1	4	8	3
Test/extra	1	0	0	0	1	0	4	24	3
Test 2	1	0	0	1	0	0	4	8	3
Test/extra	1	1	0	0	0	0	6	18	3
Test 3a	0	0	0	1	1	1	5	3	8
Test 3a	1	1	1	0	0	0	13	16	8
Test 3b	0	0	1	1	0	1	6	3	8
Test 3b	1	1	0	0	1	0	11	24	8
Test 3c	0	1	0	1	0	1	7	3	8
Test 3c	1	0	1	0	1	0	10	16	8
Test 3d	0	1	1	0	0	1	8	4	8
Test 3d	1	0	0	1	1	0	9	24	8
Test 3e	0	1	0	0	1	1	7	9	8
Test 3e	1	0	1	1	0	0	10	27	8
Test/extra	0	1	0	1	1	1	14	13	25
Test 4	0	1	1	0	1	1	16	9	25
Test 4	0	1	1	1	1	0	18	9	25
Test 4	1	1	0	1	0	1	21	8	25
Test 4	1	1	1	0	1	0	26	9	25
Test/extra	1	1	1	1	1	0	55	25	37
Test/extra	1	1	1	1	1	1	100	37	37

Note. Test/training indicates the eight objects that constituted the training set in the two conditions with three predictive cues in Study 2. These eight objects also appeared in the respective test sets. Test 2 denotes objects with a cue sum of 2, Test 3 objects with a cue sum of 3, where pairs with same letters indicate opposite cue profiles, and Test 4 objects with a cue sum of 4. Test/extra indicates objects that were additionally included in the test set to increase the differences in model predictions.

Appendix B

Accuracies of the Regression Model and the Standard Exemplar Model

In the following we report the performance of the regression model and the standard exemplar model in predicting participants' estimations. For Study 2 we additionally tested a regression model and a standard exemplar model with a free attention parameter for every cue. The predictions of the regression model were obtained by running a multiple linear regression with the cues of the training phase as predictors and participants' estimations in the last four blocks as the dependent variable. On the basis of the obtained cue weights, predictions for the test phase were made. The standard exemplar model was fitted in the same way as the simplified exemplar model, but s was allowed to vary freely for each cue. The regression model and the standard exemplar model performed worse than the mapping model and the simplified exemplar model when all cues were predictive. Only when half of the cues were predictive and cue directions were known was the mapping model better than the regression model and the standard exemplar model (both $Z = -3.92$, $p < .01$), but the regression model outperformed the simplified exemplar model ($Z = -2.95$, $p < .01$). In the condition of Study 2, where the cue directions were unknown and only three cues were predictive, the regression model ($Z = -2.95$, $p < .01$) and the standard exemplar model ($Z = -2.24$, $p = .02$) performed better than the mapping model and equally as well as the exemplar model, $Z_{\text{regression}} = -1.53$, $p = .13$; $Z_{\text{exemplar}} = -1.57$, $p = .12$. The standard exemplar model predicted the estimations of 5 (25%) participants best, the regression model provided the best estimations for 3 (15%), and the simplified exemplar model for 12 (60%). An overview of the accuracies of the regression model and the standard exemplar model in Study 2 is reported in Table B1.

Table B1: Accuracies of the regression model and the standard exemplar model in predicting participants' estimations

	Number of predictive cues							
	Six predictive cues				Three predictive cues			
	Cue directions				Cue directions			
	Known		Unknown		Known		Unknown	
	Regression	Exemplar	Regression	Exemplar	Regression	Exemplar	Regression	Exemplar
Training set								
RMSD	12.89	1.20	14.70	7.31	2.92	2.45	4.24	4.53
SD_{RMSD}	.30	1.72	2.54	6.96	1.04	1.69	1.69	2.58
r^2	.88	1.00	.81	.91	.94	.95	.85	.85
SD_{r^2}	.003	.01	.12	.14	.05	.06	.13	.14
Test set: Old								
RMSD	13.33	3.19	15.61	8.35	3.56	3.07	5.63	5.35
SD_{RMSD}	.92	3.82	3.04	6.30	1.90	2.35	2.72	2.49
r^2	.86	.98	.79	.90	.91	.92	.76	.77
SD_{r^2}	.05	.06	.11	.15	.10	.11	.23	.20
Test set: New								
RMSD	27.03	29.94	27.94	28.18	13.90	16.57	10.01	10.05
SD_{RMSD}	2.57	8.20	3.52	6.68	3.57	3.30	2.54	3.10
r^2	.39	.35	.24	.33	.36	.14	.39	.33
SD_{r^2}	.15	.18	.14	.20	.09	.11	.22	.27
Test set: Total								
RMSD	24.16	25.75	25.28	24.72	12.08	14.31	9.14	9.17
SD_{RMSD}	2.15	7.10	3.10	5.72	3.06	2.88	2.22	2.49
r^2	.37	.43	.33	.43	.45	.25	.47	.42
SD_{r^2}	.09	.15	.11	.16	.10	.12	.21	.24

Note. $N_{Total} = 80$, with $N = 20$ in each condition. *RMSD* = root mean squared deviation

Footnotes

1. Although this pattern holds true for a wide range of parameter values it should be noted that the strength of the qualitative differences in model predictions depends on the composition of the training set as well as on the parameter values for the cues. Therefore we selected training set–test set combinations where strong qualitative results should be expected.

2. For the model comparison we used the nonparametric Wilcoxon test. The standard exemplar model performed significantly worse than the mapping model in the test phase and worse than the simplified exemplar model in both conditions, in all cases $Z < -4.3$, and $p < .01$. The regression model was significantly worse than the mapping model ($Z = -5.44$, $p < .01$, for both conditions) but performed equally as well as the simplified exemplar model (in all cases $Z = -1.02$, $p = .32$).

3. We also tested a version of the mapping model that included only the three cues that were substantially correlated with the criterion. However, overall this model did not perform better than a mapping model that considered all cues.

Authors' Note

Bettina von Helversen and Jörg Rieskamp, Max Planck Institute for Human Development, Berlin, Germany. We would like to thank Anita Todd for editing a draft of this manuscript. This work has been supported by a doctoral fellowship of the International Max Planck Research School LIFE to the first author. Correspondence concerning this article should be addressed to Bettina von Helversen.

Bettina von Helversen
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany
Phone: ++49-30-82406699
Fax: ++49-30-82406394
Email: vhelvers@mpib-berlin.mpg.de

Chapter 3:
Predicting Sentencing for Low-Level Crimes:
A Cognitive Modeling Approach

Abstract

Laws and guidelines regulating legal decision making are often imposed without taking the cognitive processes of the legal decision maker into account. In the case of sentencing, this raises the question of to what extent the sentencing decisions of prosecutors and judges are consistent with legal policy. Especially in handling low-level crimes, legal personnel suffer from high case loads and time pressure, which can make it difficult to comply with the often complex rulings of the law. To understand sentencing decisions it is beneficial to consider the cognitive processes underlying the decision. An analysis of fining and incarceration decisions in cases of larceny, fraud, and forgery showed that prosecutors' sentence recommendations were not consistent with legal policy. Instead they were well described by a cognitive theory of quantitative estimation that assumes sentence recommendations rely on a categorization of cases based on their characteristics.

Predicting Sentencing for Low-Level Crimes: A Cognitive Modeling Approach

How are criminal sentences determined? Although legal systems differ from country to country, judges worldwide struggle with the problem of determining which factors should be considered and how they should be combined to form appropriate and just sentences. Even if the legal system provides guidelines to regulate the sentencing process, the question still remains how well judges and other legal personnel follow the prescribed policies (Ruback & Wroblewski, 2001). Research on sentencing has a long tradition of identifying deviations from legal policy: Extralegal factors such as race or gender have been found to influence sentencing, and in some cases legal factors are not properly taken into account (e.g., Davis, Severy, Kraus, & Whitaker, 1993; Ebbesen & Konečni, 1975; ForsterLee, ForsterLee, Horowitz, & King, 2006; Henning & Feder, 2005; Johnson, 2006; Ojmarrh, 2005). This indicates that the cognitive processes of legal professionals do not always lead to sentencing that is consistent with the sentencing policy specified by the law (Dhami & Ayton, 2001; Ebbesen & Konečni, 1975; Hertwig, 2006; Tata 1997; Van Duye, 1987).

The goal of this article is to investigate to what extent sentencing decisions deviate from legal regulations and how these deviations can be explained by cognitive models of the sentencing process. For this purpose we test whether prosecutors' sentence recommendations can be better explained by a cognitive model or by adherence to legal policy. Additionally we examine whether the same cognitive processes underlie both fining and incarceration decisions.

Heuristics in Legal Decision Making

The legal decision environment is highly complex and the workload of legal personnel heavy; decisions need to be made under time pressure and often little or no feedback regarding the quality of the decision is available (Gigerenzer, 2006). Even if specific rules exist to guide the decision process, they are often too complex to be executed in the allotted time (Ruback & Wroblewski, 2001). Not surprisingly, then, research on sentencing has found that often only a small part of the available information is used (Ebbesen & Konečni, 1975, 1981) to determine a sentence.

Heuristics are simple strategies that allow decisions to be made without much information or complex computations. Although there is disagreement on to what extent heuristics allow good decisions and how they should be formalized (Gigerenzer, 1996;

Kahneman &Tversky, 1996), there is converging evidence that heuristics provide good accounts of people's decision processes (e.g., Bröder & Schiffer, 2003; Payne, Bettman & Johnson, 1993; Rieskamp & Otto, 2006). In particular, when making complex decisions under time pressure, reliance on heuristics increases (Payne, Bettman, & Johnson, 1988; Rieskamp & Hoffrage, in press), making the legal domain an area conducive to decision-making heuristics. In fact, reliance on heuristics has been shown in several areas of legal decision making, such as bail decisions (Dhami & Ayton, 2001; Dhami, 2003; Leiser & Pachman, 2007), tort law (Guthrie, Rachlinksi, & Wistrich, 2001), and sentencing (Englich, Mussweiler & Strack, 2006; for an overview see Colwell, 2005; Engel & Gigerenzer, 2006).

Especially for the domain of low-level offenses where the decision situation can be relatively transparent and the costs of wrong decisions low, reliance on heuristics might be a way to deal with the immense workload involved. Although regrettably widely ignored by research (for an exception, see Albrecht, 1980), the majority of the cases in courts are low-level crimes and petty offenses. For example, in Germany, about 80% of the cases are punished with a fine (Langer, 1994; Meier, 2001), an alternative to incarceration that can only be imposed in minor cases. Thus, particularly in cases sentenced with a fine, heuristics might be prevalent.

Sentencing Decisions by the Prosecution

As in most legal systems, in Germany the sentence is determined by the judge. However, the judge makes this decision after hearing sentencing recommendations from both the prosecution and the defense. Research has shown that the sentencing recommendation of the prosecution is the single most important factor influencing the decision of the judge (Ebbesen & Konečni, 1975; Schünemann, 1988). For instance, Englich and Mussweiler (2001) found that, all things being equal, the recommendation of the prosecution significantly influenced a criminal's sentence; similarly Dhami and Ayton (2001) showed that in bail decisions, British magistrates followed almost without exception the recommendation of the prosecution. Additionally the prosecution can directly impose fines by penalty order. If the defendant accepts the fine, the case never goes to trial. These findings indicate that to understand which factors influence a sentence's magnitude, it is indispensable to first investigate the process by which the prosecution determines the sentence recommendation.

How should the prosecution do this? German sentencing is regulated by the German penal code (*Strafgesetzbuch*, *StGB*; Tröndle & Fischer, 2007), more specifically by articles 21, 23, 46, 47, and 49 and by decisions of the German Federal Court of Justice. Both judge and prosecution are bound by the same legal regulations. The general goal is to achieve an appropriate sentence that is proportional to the guilt of the offender. For each offense there exists a sentencing range that establishes a minimum and a maximum sentence that can be imposed. Within these often rather broad sentencing ranges, the placement of the sentence depends on the seriousness of the case and is largely left to the discretion of the judge. The judge's task, as well as the prosecution's, is to evaluate the factors mitigating or aggravating the guilt of the offender and to determine the sentence accordingly. Which factors should be considered as mitigating or aggravating is specified in the penal code. Article 46 of the *StGB* alone lists over 20 factors relevant for the sentencing decision although it cautions that it is not an exhaustive list.[1]

What the German penal code (§ 46) does not provide is explicit guidelines on how the factors should be combined. However, the German Federal Court of Justice recommends that mitigating and aggravating factors be balanced in an integrative evaluation of the overall picture (Schäfer, 2001). According to the predominant opinion in the legal literature, this is best accomplished with a three-step sentencing process: All relevant factors are evaluated according to the direction of their effect on the sentence (aggravating or mitigating), then weighted by their importance, and finally added up to form the sentence (Bruns 1985, 1988; Foth, 1985; Schäfer, 2001; but see Mösl, 1981, 1983; and Theune, 1985a, 1985b). Thus, the legal prescription asks for a linear additive decision process.

Models of Sentence Magnitude

How can the underlying cognitive process of sentencing decisions be described? In many areas of psychology multiple linear regression models are applied to analyze decision policies (Doherty & Kurz, 1996; Brehmer, 1994, Cooksey, 1996). Likewise, in the legal domain these have been the predominant models used to analyze sentencing policies and to identify which factors influence sentence magnitude (Engen & Gainy, 2000; Johnson, 2006; Kautt, 2002; Kautt & Spohn, 2002). Regression models are especially attractive to model sentencing, as the three-step model is consistent with their linear additive approach (Brehmer, 1994; Hammond, 1996). More specifically, regression models assume that quantitative judgments, such as determining the magnitude of a sentence, can be modeled as

a process of weighting and adding information (Doherty & Brehmer, 1997; Einhorn, Kleinmuntz & Kleinmuntz, 1979; Juslin, Karlsson, & Olsson, in press). Each factor is weighted according to its importance and the judgment is determined as the sum of the weighted factor values. The weights that best characterize the sentencing process are found by minimizing the squared deviation between the actual and the estimated sentence (cf. Cohen, Cohen, West, & Aiken, 2003; Cooksey, 1996):

$$(1)\ \hat{y}_p = \sum_{j=0}^{J} \beta_j c_j + \beta_0 ,$$

where the estimate, $\hat{y}_p$, for the case p is given by the sum of the product of the factor values, c_j, of the factors j with their respective weights, β_j, plus an intercept, β_0.

If prosecutors and judges in fact weigh mitigating and aggravating factors against each other and then add up the weighted factor values to arrive at a final sentence, sentencing should be well captured by multiple regression. In this case multiple regression allows us to identify the factors that influenced the sentencing decision. Furthermore, if the sentencing policy corresponds to the law, all legally relevant factors should make a significant contribution, whereas extralegal factors should not be considered. Thus, analyzing sentencing with a multiple linear regression approach allows us to compare the judges' and prosecutors' sentencing policies to the policy required by law.

The Mapping Model: A Cognitive Theory of Quantitative Estimation

Even though multiple regression can capture decision outcomes, its value as a model of human judgment processes is debatable. Researchers have doubted that people actually perform the relatively complex calculations required by multiple regression and therefore have argued that multiple regression does not provide a valid description of the cognitive process underlying a decision (Brehmer, 1994; Einhorn et al., 1979; Gigerenzer & Todd, 1999; Hoffman, 1960). In response to this criticism we have proposed an alternative, called the mapping model, that we consider to be a psychologically plausible alternative to multiple regression. The mapping model provides a cognitive theory for quantitative judgments and has been successful in predicting people's estimations (Helversen & Rieskamp, in press).

Generally, the mapping model assumes that when people make a judgment about a case or object, they assign the object to a category and use a typical criterion value for this category as an estimate. Categories are formed on the basis of previously encountered objects, and the category membership is defined by the objects' characteristics or features.

The typical criterion value of a category is represented by the median criterion value of all cases belonging to this category. For example, to estimate the selling price of a house, the mapping model assumes that one would consider the house's features that speak in favor of a high price (e.g., great location, a deck, a swimming pool), categorize the house according to its average value on theses features into a certain price class, and estimate a price that is typical for houses within this price class, that is, the median price for which houses in this category were sold for.

Helversen and Rieskamp (in press) showed that the mapping model in comparison to multiple regression was particularly suitable for predicting people's estimations if the cases' criterion values followed a skewed distribution, which is typical of sentencing decisions (Meier, 2001). Helversen and Rieskamp (in press) tested the mapping model under highly controlled experimental settings; yet these conditions are similar to the conditions of sentencing decisions, suggesting that the mapping model might be a good model for sentencing decisions.

How is the mapping model applied for sentencing decisions? Commonly, each case is described by several characteristics or factors relevant for sentencing. To apply the mapping model, first cases are categorized according to their mean value on these factors.[2] To allow comparisons of factors with different dispersions all factors are normalized by applying range frequency theory (Parducci, 1974). Using range frequency theory for normalization instead of a purely statistical technique (i.e., *z*-transformation) has the advantage that a psychologically more plausible representation of how the magnitude of a factor value is subjectively perceived by an individual is accomplished (for details see Appendix A). After normalizing all factors the mean factor value for all encountered cases is determined. This mean value represents the seriousness of the case. Next, the minimum and the maximum value of cases' seriousness are determined and the range is divided into seven equally sized categories; that is, category boundaries are chosen so that the distance between category boundaries is the same for all categories. Due to humans' limited cognitive capacities (see also Miller, 1956) only a limited number of categories is assumed. Next, the typical sentence for each category is computed by taking the median sentence of all previously encountered cases that fall into the same category. The sentence for a new case is simply determined by establishing its category membership and then using the typical sentence of that category as a sentence for the new case. Figure 7 gives an overview of the processing steps assumed by the mapping model.

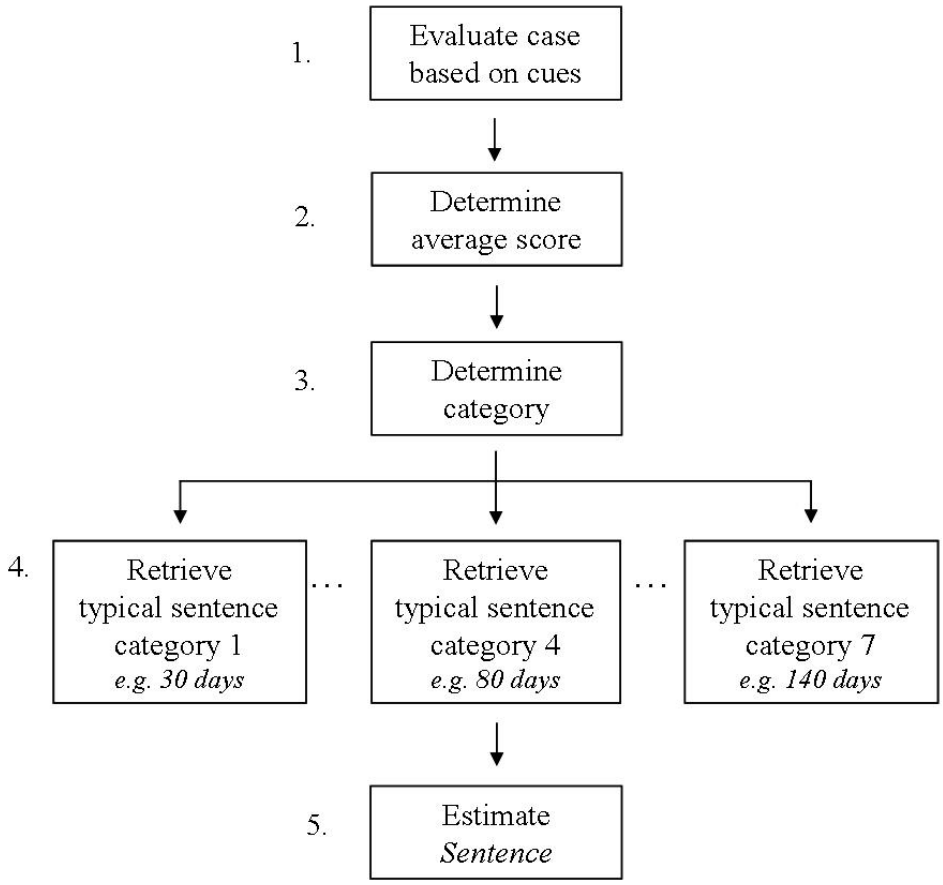

Figure 7: The processing steps of the mapping model. In the first step, the relevant cues are evaluated and rated according to their severity. In the second step the cues are integrated by establishing the average severity score. Then, the case is categorized according to its average score and the typical criterion value, that is the sentence for this category is retrieved. In the last step the retrieved criterion value is used as an estimate

To give an example how the mapping model can be applied to sentencing we will describe how a typical case would be sentenced according to the mapping model. Imagine a case of shoplifting: The defendant has confessed stealing minor goods in five cases. The net worth of the stolen goods amounts to $100 and the defendant has three prior convictions for theft. In the first step the prosecutor considers the cues, that is the characteristics of the case relevant for sentencing such as the number of charges, the amount of money stolen and the number of prior convictions. Next, she rates the severity of each cue; for instance, the amount of money stolen was low, but three prior convictions are of medium severity and so forth, thereby standardizing the cues. After that she forms an overall impression of the case,

taking the average of the cues' severity scores. Based on this average score she categorizes the case as a theft of medium seriousness and retrieves the typical sentence for this category. In the last step she determines the sentence recommendation based on the retrieved category value.

When comparing the mapping model with the regression model two important differences can be emphasized. First, unlike the regression model, the mapping model gives all factors the same weight for assigning a case to a category. Second, in contrast to the regression model, the influence of one single factor in the mapping model can interact with the other factors. The factors determine which category is used to make an estimate. Thus, how an estimate changes when one factor provides positive compared to negative evidence depends on the evidence of the other factors. Here the mapping model differs substantially from the regression model, where a factor's impact on the estimate is independent of the other factors.

Fines versus Incarceration

The second goal of this article was to investigate differences between fines and incarceration sentences. Low-level offenses can be sentenced by one or the other. Although much research has examined which factors influence sentencing length in incarceration sentences (ForsterLee et al., 2006; Johnson, 2006; Langer, 1994; Oswald, 1994; Schünemann, 1988), to our knowledge there is a lack of research on fining and the differences between incarceration and fining decisions (for an exception see Albrecht, 1980; Oswald, 1994). However, fining and incarceration are often viewed as serving different sentencing goals (Schäfer, 2001). This suggests that fining and incarceration decisions could be based on different factors and the cognitive processes underlying the decisions could also differ.

Fining decisions could be especially likely to induce heuristic decision making. As fines constitute the majority of the sentences (Meier, 2001), they represent the biggest proportion of the prosecution's workload. More serious cases might be allotted more time and be processed more systematically, as they are less frequent, incur more public interest, and have a higher probability of appeal. Thus cases sentenced by a fine could differ systematically from cases that are sentenced with incarceration. To investigate these questions, we conducted an analysis of trial records for three common offenses.

Study: Analysis of Trial Records

The first goal of the study was to model sentencing decisions for common minor offenses, investigating how well the sentencing procedure corresponds with the legal policy and if sentencing decisions are best described by a cognitive theory of quantitative estimations. The study's second goal was to examine whether fining differs systematically from incarceration decisions and which factors influence sentence magnitude in the two decisions.

We approached these goals by conducting an analysis of trial records. In comparison to an experimental approach, this type of analysis has the advantage that it is based on real cases and does not need to be limited to the small number of factors that can be manipulated in an experimental study. Furthermore, the complexity of the real cases as well as the time pressure of the daily case load could be decisive for the cognitive process underlying the sentencing decisions, favoring the analysis of real case data.

Method

We focused on three common offenses against property, namely, theft, fraud, and forgery. This allowed us to include different offenses while measuring the severity of the offense on a common scale—money—and keeping the sentencing range equal (0–5 years for a common case and 3–6 months to 10 years for an aggravated case).To investigate the sentencing process we collected trial records from a small Brandenburg Court (the *Amtsgericht Bad Freienwalde*), for the years 2003 to 2005. All records with a main charge of theft, forgery, or fraud (§§ 242, 243, 244, 248, 263, and 267) were included in the analysis. Trial records included the indictment, the transcript of the trial, orders by the prosecution, and the verdict. Based on these documents we identified offense and offender characteristics relevant for sentencing, the sentencing range, and the recommendations of the prosecution and the defense.

Categorization system. Offense and offender characteristics were classified by a categorization system that was based on the German penal code (§§ 46, 47, 52, 53, 242, 243, 244, 248, 263, and 267) in close cooperation with legal experts in the area of sentencing. Classification of a factor rested upon the indictment, the trial transcripts, and the verdict. Besides the legal factors, the categorization system also included extralegal factors that have been found to affect sentencing (e.g., Ebbesen & Konečni, 1975; ForsterLee et al., 2006). Table 12 provides an overview and a description of the factors.

Table 12: Overview of the categorization system

Factors	Description	Values
	Offender information	
Gender	Male vs. female	0 vs. 1
Nationality	German vs. non-German	0 vs. 1
Age		20–80 years
Family status	Married or single with kids vs. single and no kids	0 vs. 1
Occupational status	Employed, apprenticed, or student vs. unemployed	0 vs. 1
Economic status	Above poverty line vs. below poverty line (ca. €900 per month)	0 vs. 1
Diminished capacity	No diminished capacity vs. diminished capacity (Diminished capacity was assumed if the defendant had a psychological or medical diagnosis of a mental or organic disorder)	0 vs. 1
No. of prior convictions		0–14
Type of last sentence	Fine, incarceration, or incarceration with probation	Dummy coded
Probation status	Offender was not on probation when the offense was committed vs. was on probation	0 vs. 1
	Offense characteristics	
Net worth of property violated		€0–80,000
No. of charges		1–112
No. of offenders		1–3
Mitigating evidence I	Coded as a summary factor; one point was added if there was external pressure to commit the crime (e.g., an emergency situation or blackmail), the crime was a failed attempt, the offender's role was secondary, or the offender's capacity was diminished due to alcohol	0–2
Mitigating evidence II	One point was added if the offender had no prior convictions or the net worth of property violated was below €30	0–2
Remorse	Defendant showed no remorse vs. showed remorse, offered reparation or amends	0 vs. 1
Confession	Defendant did not confess vs. defendant confessed	0 vs. 1
Aggravating evidence	One point was added if any of the following conditions was fulfilled: a high number of offenses (> 5), over a long period of time (> 6 month); the offense was carefully planned; perseverance in the face of obstacles; incited others to commit the crime; used unnecessary violence	0-2
	Legal regulations	
Offense type	Theft, fraud, or forgery	Dummy coded
Summary penalty	A summary penalty was not given vs. a summary penalty was given	0 vs. 1
Penalty order	Sentencing by trial vs. sentencing by penalty order	0 vs. 1
Sentencing range	Max. sentence 5 years vs. max. sentence 10 years	0 vs. 1

The categorization system included personal information on the offender, as well as legally relevant factors concerning the offender's criminal and personal history. To capture the severity of the crime several characteristics of the offense were coded, such as the number of charges and the net worth of property violated. The presence of mitigating and aggravating factors concerning the conduct of the crime were coded in two summary factors capturing the amount of mitigating and aggravating evidence. If the description of a case in

the indictment and the trial protocols left doubt about the presence of a mitigating or aggravating factor the verdict was used as a reference. Only if the behavior in question was mentioned in the rationale of the verdict was it considered as mitigating or aggravating evidence. Additionally, the presence of a confession and mitigating behavior after the crime, such as remorse, were coded as two separate factors. A further mitigating summary factor coded whether the net worth of property violated was low enough to count as a less severe case (§ 248) and whether the offender had no prior record; these are two characteristics specifically identified by the German penal codes that mitigate the sentence regardless of the overall impact of property violated or of any prior record. Additionally we included three factors concerning legal regulations, such as, for instance, the sentence range applied. Finally, we did not include the recommendation of the defense in the analysis, because in most cases the defendant did not have a defense attorney present during the trial.

For most variables a nominal or ordinal level of measurement was assumed. Nominal variables were binary coded, indicating the presence or absence of a factor; ordinal variables were dichotomized by a median split. For the variables number of charges, offenders, and prior convictions, amount of mitigating or aggravating evidence, and net worth of property, an interval scale was assumed. Two independent raters coded the cases. The raters' agreement was satisfactory on all subjectively rated factors (r = .77, SD = .12). Non-random missing data were analyzed and missing values substituted with the mean of the variable, because no effect on the dependent variable was found and the overall number of cases was rather small.

Dependent variables. Dependent variables were the type of sentence (fine or incarceration) and the number and magnitude of daily payments (for fines) and the length of a prison term in months (for incarceration) as recommended by the prosecution and the verdict. According to the German legal system a fine is constructed as a number of daily payments of a certain magnitude. The number is determined in correspondence to the severity of the crime, whereas the magnitude depends on the income of the defendant. As the aim of this study was to compare sentencing for prison terms and fines we focused on the number of daily payments as the dependent variable for fines corresponding to length of prison sentence. The number of daily payments can vary between 5 and 365; more severe offenses are sentenced by incarceration. The dependent variable for incarceration length was number of months sentenced to prison, irrespective of whether the offender was let off with

probation. To identify the differences between fines and incarceration, we analyzed the sentences for fines and incarceration separately.

Description of the court, the offenses, and the offenders. The Amtsgericht Bad Freienwalde is a small court in the Brandenburg district of Märkisch-Oderland, close to the Polish border under the jurisdiction of the Frankfurt (Oder) district attorney's office. The city of Bad Freienwalde has a population of 13,000 with an unemployment rate of 12%. Overall, 99 cases of theft, fraud, and forgery were tried in this court during 2003 and 2004. From the 99 cases, 15 were excluded because the major charge was none of the offenses under consideration, juvenile law was applied, or the case did not lead to a conviction. Of the remaining 84 cases, 82% were tried by the same judge. The 84 cases were prosecuted by 45 different attorneys with a maximum of 5 cases by the same attorney. In 49 cases the main charge was theft, in 20 it was fraud, and in 15, forgery. On average, property worth €2,497 was violated (*SD* = €8,826). The offenders were predominantly German males; 69 were men and 15 women. Eight offenders did not have German citizenship. The mean age of the offender was 36 years, ranging from 20 to 80 years. About half of the offenders were sentenced to a fine (*M* = 48 days; *SD* = 27) and half to a prison term (*M* = 8 months; *SD* = 6).

Model selection. The main goal of our study was to identify the cognitive process underlying sentencing and to determine if a cognitive model of sentencing could predict the magnitude of a sentence. For this purpose we tested which theory describes the sentencing process better: legal policy as modeled by a multiple linear regression model (e.g., Cooksey, 1996) or the mapping model, a cognitive theory for quantitative estimation (Helversen & Rieskamp, in press).

Testing these two models on the data of real cases raised two crucial methodological problems: First, real cases involve an enormous number of factors that could potentially predict the sentence. In our cases we recorded 22 factors that could influence the sentencing decision. How can we find out which factors have a substantial effect? One common technique when using regression models for identifying important factors relies on significance tests. In these models the estimated impact of a factor can depend on the other factors included in the regression equation, so that often procedures are performed where factors are step-wise either included or excluded from the regression equation (cf., Cohen et al., 2003). However, when considering a larger number of factors this procedure is very unsatisfying, because factors that were added to the equation at the beginning of a step-wise forward procedure might not have been added had other factors already been included.

Therefore, different statistical procedures applied to the same original set of factors often lead to inconsistent results (i.e., different regression equations), which can lead to very different conclusions.

The second methodological problem we faced concerns the models' complexity, that is, their flexibility in describing different results. In particular, we were interested in testing the regression model against the mapping model; these models differ in their number of free parameters and therefore in their potential to describe different processes. Therefore, we sought a methodology that would take the models' complexity into account when testing them against each other.

To tackle these two methodological problems we followed a Bayesian approach, specifically the Bayesian model averaging (BMA) method (see Raftery, 1995, and also Raftery, Madigan, & Hoeting, 1997). This Bayesian method identifies the model or the models that are most probable given the data. Furthermore, BMA provides reliable estimates of the predictors' influence on the dependent variable and it allows comparison of models of different complexities by taking the models' free parameters into account. BMA was proposed especially to examine the uncertainty of parameter estimates and for model selection. To identify the most probable models, the Bayesian method calculates the posterior probability of a model given the observed data. Pragmatically this is performed by determining the Bayesian information criterion (BIC), which approximates the so-called Bayes factor (Raftery, 1995; Schwarz, 1978). The method additionally allows one to specify the probability that a factor will have an impact on the dependent variable: Taking model uncertainty fully into account, the average amount of evidence speaking for an effect of a factor is determined by summing the posterior probabilities of all models that include this factor (for details see Appendix B).

The most reliable method for model selection, according to Raftery (1995), is to construct all possible models that can be built with the available factors and then select the models with the highest posterior probability given the data. However, including all candidate predictor variables would result in an enormous number of possible models, as 15 predictor variables already amount to 32,768 models. Thus we reduced the number of factors by first including all factors that substantially correlated with the dependent variable (i.e., showed a value of $r > .3$) and then additionally adding factors such as confession or remorse that were not necessarily correlated with sentence magnitude but are of special theoretical importance, because they frequently appear as mitigating reasons in the rationale of the

verdict. For the fines we included 11 factors and for the incarceration decisions 9 factors (see Tables 2 and 3).

Next we calculated the BIC values for all models resulting from all possible combinations of the factors. This amounted to 2,048 models in the case of fining decisions and 512 models in the case of incarceration decisions for each model class, the mapping models and the regression models. We first ran the analysis separately for the two model classes, to investigate if the factors identified by the two types of models would differ. Then we included all models in the comparison to identify which class of models most probably underlies the decision behavior given the observed data.

For all of the models we calculated the BIC′ value based on the amount of variance explained (R^2) as a measure of goodness-of-fit of the model and the number of free parameters (see Raftery, 1995). Details on the computation and the equations can be found in Appendix B. The BIC′ value gives the odds with which a specific model is preferred to a baseline. In the case of regression, usually a null model is chosen as a baseline model. The null model only includes an intercept (i.e., estimates the mean criterion value for all objects) and no predictor (i.e., free parameter). It explains zero of the variance in the data and its BIC′ is zero (see Equation 2); The BIC_k' of a specific model M_k is defined so that if the BIC_k' value is positive, the null model is preferred, while a negative BIC_k' value provides evidence for the model M_k under consideration. The lower the BIC_k' value, the more the model is supported by the data.

(2) $$BIC_k' = n \log\left(1 - R_k^2\right) + q_k \log n$$

where R_k^2 is the value of R^2 for model M_k, q_k is the number of free parameters for that model, and n is the number of data points.

For the regression models a least squares regression was run with the factors as predictor variables and the sentence recommendation of the prosecution as the dependent variable. For the mapping models the category borders and the typical sentence for each category were estimated from the data. First the perceived factor score was calculated based on range frequency theory with one free parameter for all factors, capturing the relative importance of range and frequency information (for details see Appendix A). Then case seriousness was computed by averaging the factor scores over all factors, the minimum and maximum case seriousness was determined, and the range was divided into seven equally sized categories. For each category the typical sentence was calculated by taking the median

of all cases that fell into this category. The typical sentence was estimated for all cases falling into one category and the amount of variance in the sentence recommendation of the prosecution explained (R^2) was computed. Based on the BIC′ value we calculated the posterior probability of each model, assuming equal priors for all models. Additionally we computed the probability of each factor being included and an approximation of a Bayesian point estimator of beta weights and standard errors for each factor (see Appendix B).

Results

Overall, the more parsimonious mapping models offered the more probable description of the data, but both model types identified the same factors as influencing sentencing. Although sentencing decisions for fines and prison times were both based on the factors net worth of property and number of charges, the role of mitigating and aggravating evidence differed for the two sentence types. Fining decisions were more influenced by aggravating evidence and the number of prior convictions while incarceration length was more affected by mitigating evidence (II). Neither for fines nor for incarceration decisions did extralegal factors such as sex, age, or nationality play a role. In the following we report the results of the analysis for fines and incarceration separately.

Magnitude of fines. Overall, the verdict could be almost perfectly predicted by the recommendation of the prosecution (r = .99), as illustrated in Figure 8.

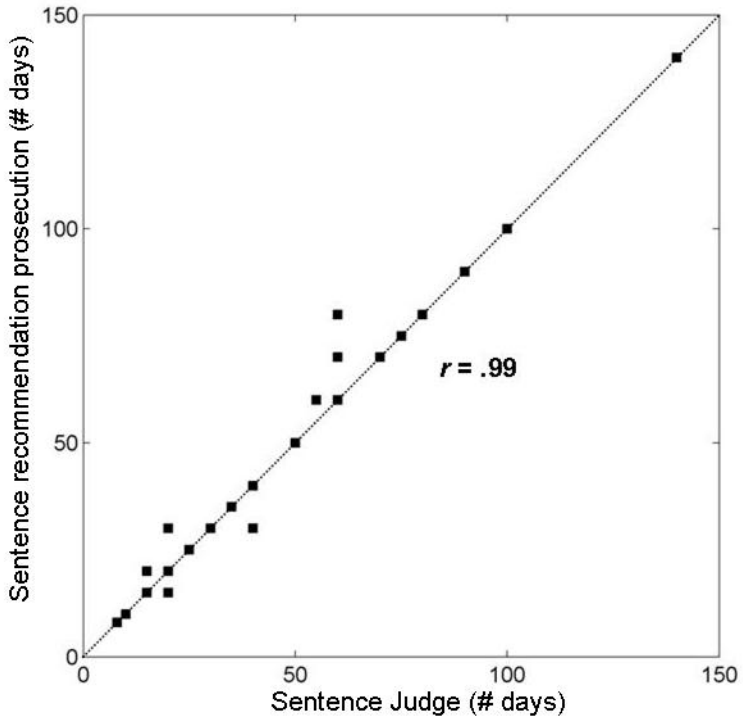

Figure 8: Scatter plot of the sentence recommendation for fines by the prosecution and the corresponding verdict by the judge. The magnitude of the fines is given in number of days a payment has to be made.

Accordingly, we concentrated on the recommendation of the prosecution as the more interesting dependent variable. The recommended sentence, in turn, correlated significantly with a number of offense and offender characteristics (see Table 13). As expected, the presence of a confession and mitigating evidence II, coding low worth of property violated and no prior record, correlated negatively with the recommended sentence. The net worth of the property violated, the number of prior convictions, the number of charges, and the amount of aggravating evidence correlated positively with the magnitude of the sentence. All other factors did not correlate significantly with sentence length or showed no variation in the sample.

Table 13: Results of correlation analysis and model comparison for fines

	Fines (no. of days) Pearson correlation (*p* values)	Five best models					Probability	Beta	*SD*
		Mapping model 1	Mapping model 2	Mapping model 3	Mapping model 4	Mapping model 5			
Age	.34 (.02)	○	○	○	○	○	.08	.03	.11
No. of prior convictions	.32 (.03)	●	○	●	○	○	.56	.19	.10
No. of charges	.36 (.02)	●	●	○	●	○	.64	.16	.17
Net worth of property	.46 (.001)	●	●	●	●	●	.97	.36	.11
Confession	-.50 (.001)	○	○	○	○	○	.15	-.08	.19
Penalty order	.50 (.001)	●	●	●	●	●	.91	.53	.14
Summary penalty	.53 (.001)	○	○	●	○	●	.46	.24	.12
Aggravating evidence	.39 (.01)	●	○	●	○	○	.59	.32	.13
Mitigating evidence II	-.48 (.001)	○	●	○	○	○	.25	.15	.12
Remorse	-.20 (.20)	○	○	○	●	○	.13	-.16	.11
Nationality	.32 (-03)	○	○	○	○	○	.14	.08	.12
PMP		.15	.13	.09	.07	.06			
BIC′		-56	-55	-55	-54	-54			
R^2		.74	.74	.73	.73	.73			

Note: N= 44; Probability denotes the probability that the factor had an effect and is given by Equation B3 (Appendix B). BIC′ denotes the Bayesian Information Criterion. PMP denotes posterior model probability. An open circle denotes that a factor was not included in the model; a solid circle denotes that a factor is included in the model. For the analyses, the

factors confession, remorse, and mitigating evidence II were recoded so that they correlated positively with sentence magnitude. The five best models all belonged to the class of mapping models

Modeling—critical factors. Model analysis showed that a few factors are sufficient to describe the data. BMA for the two model classes gave a similar picture of which factors influence sentencing. Altogether 11 factors were considered (see Table 13), resulting in 2,048 possible models for each model class. Thus the prior probability of a model was about .0005. Of the 2,048 linear regression models under evaluation, 95% had a posterior probability below .002, with the two best models reaching a posterior probability of 5% and 6% and explaining 64% and 68% of the variance in the sentencing recommendations, respectively. There was strong evidence that the factors net worth of property, penalty order, and aggravating evidence affected sentence recommendation. Additionally there was weak evidence for the factors summary penalty and number of prior convictions. The estimated beta weights can be found in Table 2.

Applying the BMA method to the class of mapping models similarly led to discarding a large proportion of models: 96% had a posterior probability below .001. However, the two best models reached a posterior probability of 15% and 13%, respectively. They both explained a much higher amount of variance (74%) in the sentence recommendations than the best regression models. Similar to the regression models, there was strong evidence for the factors net worth of property and penalty order. The factors aggravating evidence, number of prior convictions, and number of charges were supported by some evidence. In contrast to the regression model the factor summary penalty received less support.

In sum, the BMA analyses of the two model classes rendered that the choice of model had only a slight influence on which factors were identified as important. In both model classes, the most important factors were net worth of property, whether the sentence was recommended by a penalty order or after a trial, and the presence of aggravating evidence. Additionally the number of prior convictions, the number of charges, and if the sentence was a summary penalty played a role, while age, nationality, and a confession or other mitigating evidence did not influence the sentence recommendation. This is clearly inconsistent with the legal requirement that all legally relevant factors be taken into account. Particularly surprising is that confession and remorse were not considered, as they are usually mentioned as extenuating factors in the rationale for the verdict.

Model comparison. After examining which factors influenced the sentencing decision in cases punished with a fine, we now tested which type of model was better suited to explain the decision process underlying fining, mapping or regression. For this comparison we included all models and calculated the posterior probabilities, assuming that all models have the same prior probability. This resulted in a comparison of 4,096 models with a prior probability of .0002. Over all models, 17 reached a posterior probability above .01, summing up to a joint probability of .74, compared with a joint probability of .26 for the remaining 4,079 models. All of them belonged to the class of mapping models (see Table 2 for the five best models). Overall, the mapping models reached a much higher posterior probability: The joint posterior probability of all mapping models was .99999 compared to .00001 for the regression models. This is illustrated by Figure 9, showing the posterior probabilities of the best 1,500 models. The majority clearly belong to the mapping model class.

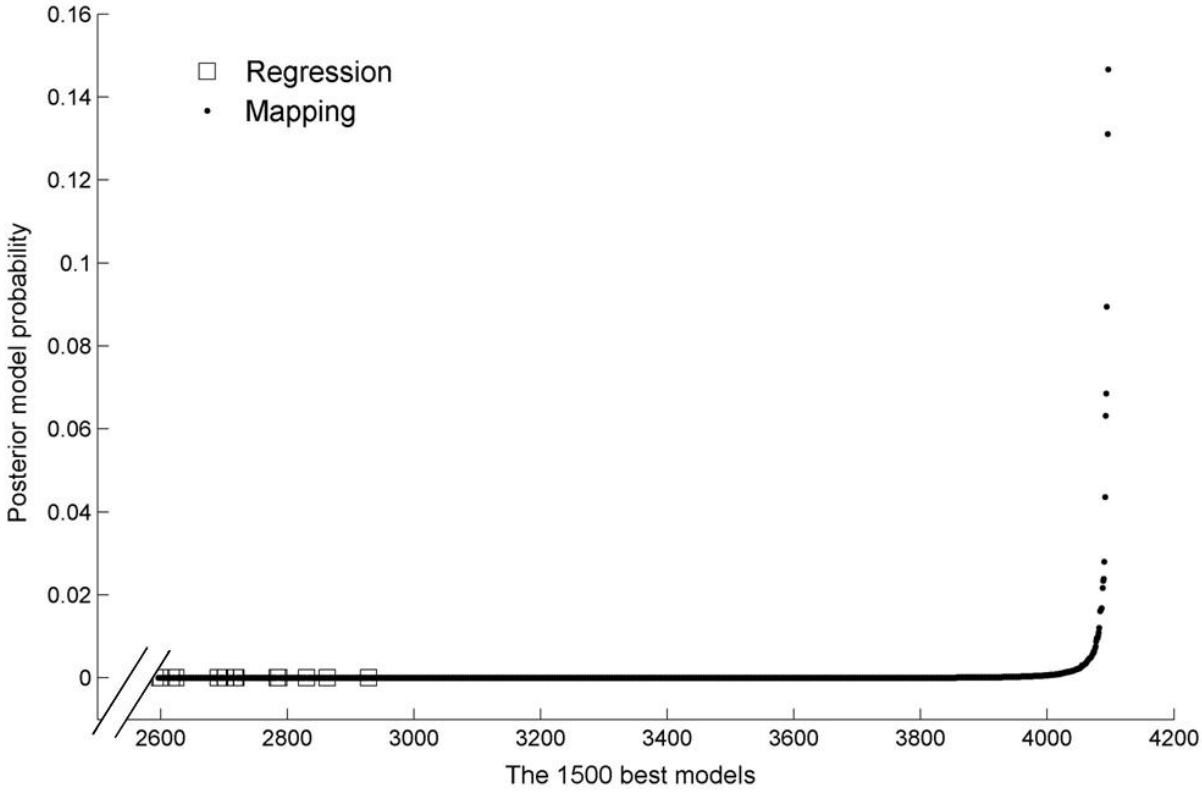

Figure 9: The posterior model probability of the best 1,500 of all 4,096 models to describe the fining process, differentiated by model class. Of the 1,500 best models, 99% belong to the class of mapping models and 1% to the class of regression models.

Incarceration length. Similar to the fines, the recommendation of the prosecution was the best predictor of sentence length (r = .95). Accordingly, we again focused on the sentence recommendation of the prosecution as the main dependent variable. Altogether we considered nine factors. Seven offense or offender characteristics correlated above .3 with the length of prison sentence. As expected, the factors net worth of property violated, summary penalty, aggravating evidence, number of charges, and number of offenders correlated positively with the sentence length, while the second mitigating factor (coding a low worth of property violated and no prior record) correlated negatively with recommended sentence length (see Table 14). The factor penalty order was not applicable as a sentence by penalty order is not allowed for prison sentences. Somewhat unexpectedly, the presence of a confession and special circumstances leading to diminished capacity correlated positively with sentence length. This effect, however, is probably due to the comparatively serious nature of these cases and does not reflect a negative evaluation of these factors for sentencing. Although remorse did not correlate with sentence length, we additionally included it in the analysis.

Table 14: Results of correlation analysis and model comparisons for incarceration

	Incarceration (no. of months) Pearson correlation (p value)	Five best models					Probability	Beta	*SD*
		Mapping model 1	Mapping model 2	Mapping model 3	Mapping model 4	Mapping model 5			
No. of charges	.40 (.01)	○	○	●	●	●	.32	.31	.01
Diminished capacity	.41 (.01)	○	●	●	●	●	.45	.41	.01
Net worth of property	.62 (.001)	●	●	●	●	○	.91	.57	.01
Summary penalty	.65 (.001)	●	○	○	○	○	.61	.20	.02
Aggravating evidence	.58 (.001)	●	○	○	○	●	.63	-.10	.02
Mitigating evidence II	-.41 (.01)	●	●	○	●	○	.78	.24	.01
No. of offender	.31(.05)	○	○	○	○	○	.04	-.01	.01
Confession	.29 (.07)	○	○	○	○	○	.03	-.09	.01
Remorse	-.01 (.98)	●	○	○	○	○	.58	.07	.01
PMP		.53	.10	.10	.07	.06			
BIC′		-49	-46	-46	-45	-45			
R^2		.82	.76	.76	.78	.75			

Note: N = 40; Probability denotes that the probability that the factor has an effect and is given by Equation B3 (Appendix B). BIC′ denotes the Bayesian Information Criterion PMP denotes the posterior model probability. A solid circle denotes that a factor was included in

the model; an open circle denotes that a factor was not included in the model. For the analyses, the factors mitigating evidence II and remorse were recoded so that they correlated positively with sentence magnitude. The five best models all belonged to the class of mapping models

Modeling—critical factors. Similarly to the analysis of the fining decisions, we used the BMA method to determine the factors with the highest probability of influencing sentence length. The factors included in the models were number of charges, net worth of property, diminished capacity, mitigating (II) and aggravating evidence, confession, summary penalty, number of offenders, and remorse, resulting in 512 models per model class with a prior probability of .002.

Five regression models reached a posterior probability above .05, with the best model clearly superior to the other models with a probability of .28, compared to the second best model with a probability of .11. The best model explained 75% of the variance in the recommended incarceration length and reached a BIC′ value of –41. There was strong evidence for the effect of the factors number of charges, net worth of property, and diminished capacity. Additionally, there was some support that mitigating evidence II influenced sentencing recommendations for prison terms. The corresponding beta weights can be found in Table 14.

For the mapping models, also five models reached a probability above .05. The best model reached a probability of .55 and explained 82% of the variance in sentence length, much more than the best regression model or the second best mapping model with a posterior probability of .10 and an r^2 of .76. However, the factors supported by the mapping models differed from the factors supported by the regression models. Similar to the regression models, net worth of property received strong support, and mitigating evidence II some support. However, there was hardly any evidence for number of charges, and diminished capacity was somewhat less important. Instead, there was additional evidence for the summary penalty, aggravating evidence, and remorse.

In sum, the analyses showed consistently that—despite the stipulations of the law—only a few factors were necessary to describe sentencing. However, which factors were considered important differed between the two model classes. Although both model classes supported the factors net worth of property violated, mitigating evidence II, and diminished capacity, applying the regression models provided evidence for the factor number of charges,

whereas the mapping models indicated the factors summary penalty, aggravating evidence, and remorse.

Model comparison. To find out which class of models was better suited to explain incarceration decisions, we again entered all models in a joint comparison. The final analysis comparing 1,024 models from both model classes supported the mapping model as the superior type of model. The best five models belonged to this model class (see Table 14). The posterior probabilities of these models added up to a joint probability of .86, compared with a probability of .14 for the remaining 1,019 models. Again, the class of mapping models was more strongly supported than the regression models. The posterior probability of all mapping models added up to .96, compared to .04 for the regression models. This is illustrated by Figure 10, depicting the posterior probabilities of the best 100 models.

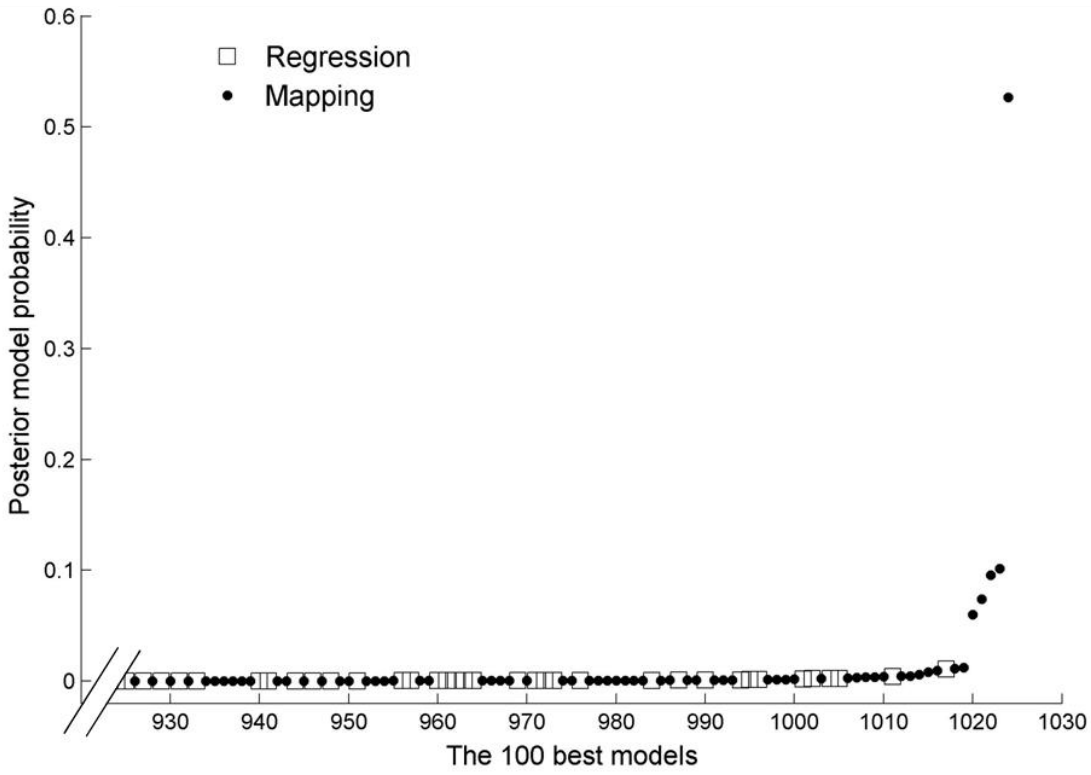

Figure 10: The posterior model probability of the best 100 models describing the incarceration decisions, differentiated by model class. Of the 100 best models, 65% belong to the class of mapping models and 35% to the class of regression models

The joint model comparison also supported the evaluation of the factors' importance by the mapping model (see Table 14). There was strong evidence for the factors net worth of

property and mitigating evidence (II), and some support for summary penalty, aggravating evidence, remorse, and diminished capacity.

Discussion

There are two ways in which sentencing decisions can deviate from the law: First, the decision can be based on a different set of factors than required by the law; second, the way these factors lead to a sentence can be inconsistent with the prescribed legal policy. The present article examined both routes by testing two different models of decision making and by identifying the crucial factors influencing sentencing, following a Bayesian approach.

The model comparison test allowed us to identify which type of model—one consistent with the legal theory or one derived from cognitive psychology—captured the sentencing decisions best. Furthermore we were able to identify the factors that were crucial for each model class to predict the sentencing. Our results show that the prosecutors neither considered all factors required by law nor exhibited decision processes consistent with the policy assumed by the legal literature. Instead, the decisions of the prosecutors were best described by a heuristic for quantitative estimation, the mapping model (Helversen & Rieskamp, in press). In the following we will first discuss the results on which factors predicted sentence recommendations and differences between fines and incarceration sentences. Then we will turn to the model comparison and the significance of the BMA method for sentencing research. Finally we will discuss limitations of the current study.

Predictors of Sentencing Decisions

Prosecutors clearly deviated from the law concerning the factors that had an impact on sentencing length. According to the law, all legally relevant factors in the analysis should have affected the sentence recommendation. However, in both types of sentencing decisions only a few factors were sufficient to predict the prosecutors' recommendations. It is in particular surprising that factors such as confession or remorse did not always lead to lower sentences, as they are usually stated as mitigating factors in the rationale for the verdict. The results are, however, in line with psychological research on judgment and decision making, which has repeatedly shown that humans often lack insight into their judgment policies (Brehmer & Brehmer, 1988) and tend to base their decisions on only a few factors (Brehmer, 1994).

Interestingly, the factors influencing fining and incarceration decisions varied substantially. For one, the magnitude of the fine was higher if the sentence was imposed via penalty order than if by trial, whereas incarceration length was influenced by diminished capacity. However, as there were no cases of diminished capacity in the sample receiving fines and sentencing by penalty order is not allowed for incarceration sentences, these differential effects are not very surprising. More interestingly, fines were influenced by prior record and aggravating evidence, but not by mitigating evidence. This suggests that the prosecution, deciding which factors were relevant, might have relied on an image of a "typical" case. Factors that indicated deviation from the norm were considered for the sentence while factors that constituted the "normal" case were not (Mösl, 1983; Tata, 1997). Fines are usually imposed in less serious cases. Thus, in cases punished by a fine the prosecution might have already "used up" the influence of any mitigating information by sparing the offender an incarceration sentence, while in cases punished with incarceration the mitigating information was taken into account, reducing sentence length.

Model Comparison

In both types of sentencing decisions, our analyses clearly illustrated that cognitively derived mapping model provided a much better explanation for the sentencing process than the regression model that is consistent with legal regulations. For the fining decisions, just about any mapping model was more probable than a regression model. Even in the incarceration decisions the five best models belonged to the mapping model class. These results are in line with those of Helversen and Rieskamp (in press), who demonstrated the success of the mapping model in comparison to the regression model in a laboratory estimation task. Because the regression model was outperformed by the mapping model, this result suggests that prosecutors do not weigh each factor individually and sum up the weighted evidence as one would expect from standard legal procedure. Instead, the cognitive process underlying sentencing decisions was more in line with the mapping model. Therefore, when prosecutors make sentencing decisions they apparently use the evidence provided to group cases of similar seriousness together, where the seriousness of a case depends on its average value on the factors considered relevant. Finally, a typical sentence is stored for each category and used to evaluate a present case.

The finding that cognitive models are more suitable to predict legal decision making is consistent with previous findings, indicating that legal decision-making processes often do

not concur with the procedures assumed by the law (e.g., Dhami & Ayton, 2001; Hertwig, 2006; Van Duyne, 1987). However, although our study illustrates that a cognitive model was more suitable to predict sentencing than a model consistent with standard legal procedure, we emphasize that following the mapping model to make sentencing decisions does not necessarily represent a case of biased decision making. In contrast, Helversen and Rieskamp (in press) showed that in situations in which the criterion is nonlinearly distributed, the mapping model was more accurate in predicting the criterion than a regression model. Thus, in sentencing situations in which the distribution of the cases' seriousness is highly skewed, the mapping model might be, in fact, more suitable than a regression model for making sentencing decisions. Particularly in low-level crimes, where legal decision makers operate under severe time constraints, making sentencing decisions according to the mapping model could be an adaptive response.

Nevertheless, making a decision according to the mapping model compared to a weighted additive model will often lead to different sentences. This raises the question of which process sentencing should follow. It also resonates with a discussion in the German legal literature in the 1980s. Instigated by a decision of the German Federal Court of Justice, the relevance of "normal" and "average" cases as reference points for sentencing was discussed (see Bruns, 1988; Mösl, 1981, 1983; Theune, 1985a, 1985b). Likewise in England, similarity-based decision aids for sentencing have been under discussion (e.g., Tata, 1998). In principle, because the German penal code does not regulate how the relevant factors should be integrated, processes as assumed by the mapping model might be legally justifiable. Although this is ultimately a legal question, psychological insights into the cognitive processes underlying legal decisions could inform a legal discussion on sentencing laws and might provide valuable input for the development of institutions.

Bayesian Approach

The way we analyzed the data and tested the two competing models differs substantially from the standard approach taken in policy-capturing research (e.g., Cooksey, 1996). According to the standard approach, one single regression model is estimated by applying a specific statistical test procedure. This approach has the disadvantage that it can lead to rather different results and conclusions depending on the statistical procedure chosen. Moreover, the interpretation of the influence of single factors is rather complicated, because the influence depends on the other factors included in the equation.

In contrast, the Bayesian approach we followed led us to consider all possible models that could be constructed with the available predictors and for each model the posterior probability was estimated. The two competing model classes were tested against each other by considering all models of each class and not simply one best model. This model comparison test provided very strong empirical support for the mapping model. Moreover, by considering which factors were included in models with large posterior probabilities, it was possible to provide more reliable conclusion about the factors that are important for sentencing decisions.

Limitations of the Study

Our study focused on one single German court. This naturally raises the question of how well the results generalize. Many studies have shown the importance of location and the legal culture of a jurisdictional district (e.g., Johnson, 2006; Kautt, 2002; Langer 1994). Especially, which factors influence sentence magnitude could differ between districts and thus our results concerning the importance of factors should be treated with caution. Furthermore, our results were based on a rather small sample, which could reduce the generalizability of the results even within the jurisdictional district. Nevertheless, for the restricted data set we could illustrate the benefits of a cognitively inspired approach to legal decision making. Future research is necessary to test if these results can be replicated with larger samples for a wider range of jurisdictional districts. Although this needs to be tested, we do not have a reason to assume that prosecutors from Brandenburg differ in their cognitive processes from prosecutors in other parts of Germany. If anything, a higher case load and more time pressure should be expected.

Even when generalizing outside of Germany, similar results might be anticipated, given that the general features of the task remain the same. That is, as long as the prosecutor or the judge has to integrate several factors to determine a final sentence, the mapping model could offer a valid description of the process. However, legal systems where sentencing is strictly regulated by sentencing guidelines, as, for instance, in the United States, could provide exceptions. Thus further studies investigating the generalizability of the utility of the mapping model to explain sentencing are necessary.

In a similar vein, it is important to note that this study focused on low-level offenses. It is an open question if the same cognitive processes underlie the sentencing of more severe cases, such as capital crimes. It appears reasonable that for more severe cases more factors

are taken into account for sentencing decisions, which therefore might be more in line with legal policy.

Conclusion and Outlook

This paper provides evidence that in sentencing, cognitive models are necessary to understand the decision process. Our results suggest that the sentence recommendations of prosecutors were not consistent with the requirements of the law; instead, sentence recommendations were well described by the mapping model, a cognitive theory for quantitative estimation (Helversen & Rieskamp, in press). This study joins a growing body of research questioning the ability of decision makers to comply with legal regulations and emphasizes the importance of understanding cognitive processes for the development of institutions.

Appendices

Appendix A

Range Frequency Theory

According to range frequency theory (Parducci, 1974), human judgments of magnitudes and size are context dependent, that is, they depend on the range of the stimulus values as well as on the frequency with which a stimulus value appears. The judged magnitude J of a stimulus i is given by the weighted sum of the range value R and the frequency value F (cf. Parducci, 1974, p. 209):

(A1) $J_i = wR_i + (1 - w)F_{i,}$

with $0 < w < 1$. The range value R represents the proportion of the current range below the current stimulus S_i:

(A2) $R_i = (S_i - S_{\min})/(S_{\max} - S_{\min})$,

where S_i denotes the current stimulus value and $S_{\min}$ and $S_{\max}$ are respectively the smallest and the largest stimulus in the set.

The frequency value F_i represents the proportion of all current values below the current stimulus:

(A3) $F_i = (r_i - 1) / (N - 1)$,

where F_i represent the frequency value of the stimulus i, r_i is the rank of stimulus $i,$ and N the number of stimuli in the set.

Appendix B

Bayesian Model Averaging

The Bayesian information criterion (BIC) gives the odds with which a specific model is preferred to a baseline model.To calculate a model's BIC′ value we compared it with the null model (a baseline model with no independent variables), following Raftery (1995, Equation 26, p. 135):

(B1) $BIC'_k = n\log\left(1 - R_k^{\,2}\right) + q_k \log n$,

where R^2_k is the value of R^2 for model M_k, q_k is the number of free parameters for that model, and n is the number of data points. The BIC'_k gives the BIC value for the null model compared to the model M_k. The BIC' of the null model is zero. Accordingly, if the BIC'_k is positive the null model is preferred to the model M_k. However, if the BIC'_k is negative, model M_k is preferred to the null model, and the smaller the BIC'_k, the more M_k is supported by the data.

The posterior probability of a model is defined as:

(B2) $p\left(M_k|D\right) \approx \exp\left(-\tfrac{1}{2} BIC'_k\right) / \sum_{l=1}^{K} \exp\left(-\tfrac{1}{2} BIC'_l\right)$,

(cf. Raftery, 1995, Equation 35, p. 145) where p gives the probability of model M_k given the data D in comparison with all models from set K assuming an equal prior probability of $1/k$ for all models.

The posterior probability pr that a factor B has an effect ($B \neq 0$) is given by the sum of the posterior probabilities of all models $p\left(M_k|D\right)$ that include B, here referred to as model set A:

(B3) $pr\left[B \neq 0|D\right] = \sum_{A} p\left(M_k|D\right)$,

(cf. Raftery, 1995, Equation 36, p. 145).

The beta weight and the standard error of the beta weights can be estimated by an approximation to a Bayesian point estimator and an analogue of the standard error. Approximations are given by:

(B4) $E\left[\beta_1|D, \beta_1 \neq 0\right] \approx \sum_{A} \hat{\beta}_1(k) p'\left(M_k|D\right)$,

(cf. Raftery, 1995, Equations 38 and 39, p. 146), where

$p'\left(M_k|D\right) = p\left(M_k|D\right) / pr\left[\beta_1 \neq 0|D\right]$, E denotes the expected value of the beta weight β_1, and $\hat{\beta}_1(k)$ is the maximum likelihood estimator of β_1 under Model M_k.

Respectively, the standard error can be approximated by:

(B5) $SD^2\left[\beta_1|D, \beta_1 \neq 0\right] \approx \sum_A \left[se_1^2(k) + \hat{\beta}_1(k)^2\right] p'\left(M_k|D\right) - E\left[\beta_1|D, \beta_1 \neq 0\right]^2$,

where $se_1^2(k)$ is the standard error of β_1 under Model M_k (cf. Raftery, 1995, p. 146).

Authors' Note

Bettina von Helversen and Jörg Rieskamp, Max Planck Institute for Human Development, Berlin, Germany. We would like to thank attorney M. Neff and the prosecution authority of Eberswalde for providing us access to the trial records. We are very grateful to Christoph Engel, Stefan Bechthold, Stefan Tontrupp, Andreas van den Eikel, and Tobias Lubitz for their help and advice in devising the categorization system. We gratefully acknowledge Patrizia Ianiro, Daria Antonenko, and Cornelia Büchling's commitment and helpful ideas in coding and analyzing the data. We would like to thank Anita Todd for editing a draft of this manuscript. This work has been supported by a doctoral fellowship of the International Max Planck Research School LIFE to the first author. Correspondence concerning this article should be addressed to Bettina von Helversen.

Bettina von Helversen
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany
Phone: (+49 30) 82406 699
Fax: (+49 03) 82406 394
Email: vhelvers@mpib-berlin.mpg.de

Footnotes

1. Besides the factors stated in § 46, German law allows sentence adjustments to achieve general prevention as well as specific prevention objectives (Meier, 2001; Schäfer, 2001). Furthermore the sentencing range can be lowered if mitigating reasons as specified in articles 21, 23, and 49, exist. As our sample did not included mitigated sentencing ranges according to these articles, we relied on the sentencing ranges as specified for common and aggravated cases of theft (§242 ff.), fraud (§263), and forgery (§267).

2. To simplify the statistical analysis we inverted all factors that were negatively correlated with sentence magnitude, so that after inversion all factors were positively correlated with sentence magnitude. Please note that this is only a statistical simplification; alternatively the difference between the mean score on aggravating factors and the mean score on mitigating factors could be taken.

General Discussion

The goal of this research was to investigate the processes underlying estimation from multiple cues and to examine if a heuristic model can be formalized to describe the estimation process. I proposed the mapping model as a possible cognitive theory for estimation and tested it in multiple experiments against several models of estimation put forward in the literature. Overall, the results clearly supported the mapping model as a realistic model of human cognition. It described participants' estimations in diverse laboratory tasks, reaching from estimating the toxicity of bugs, the probability of cure from a disease, and the evaluation score of job candidates. Furthermore, it was well suited to capture prosecutors' sentence recommendations for low-level crimes. Additionally, the research provided evidence that estimation processes are adapted to the environment. Thus how well the mapping model described participants' estimations depended on the characteristics of the task. In the following, I will discuss under which conditions the mapping model performed well, when other models might be better suited to describe the peoples' estimations and the resulting implications for the estimation process. Further, I will discuss the problems of model selection and the methods I used to address them. Finally, I will consider possible extensions and limitations of the approach and its generalizability to other areas of research.

Mapping Model

Overall, the mapping model proved to be a suitable model for human estimation. In several studies it described participants' estimations as good as or better than a linear regression model, an exemplar model, or QuickEst, a noncompensatory heuristic. Even though the mapping model is a deterministic model that ignores interindividual differences between the participants, it was surprisingly good at predicting individual estimations.

Consistent with the idea that cognitive processes are adapted to the structure of the environment, the model's success was clearly dependent on the structure of the estimation task. In my dissertation I specified conditions under which the mapping model predicts participants' estimations well. The first chapter provided evidence that the mapping model is influenced by the statistical structure of the estimation environment. It performed well in nonlinear environments, that is, if the criterion was a nonlinear function of the cues or followed a J-shaped distribution. The second chapter highlighted the importance of another

aspect of the task: the availability of explicit task knowledge, and the ease with which it can be acquired. If participants were informed about the cue directions, the mapping model was clearly the best model to predict participants' estimations. Furthermore, also if participants could easily abstract explicit knowledge about the cues during training, the mapping model outperformed the other models. Last, but not least, the mapping model did not only describe participants' estimations in a highly constrained experimental task but the laboratory results shown in the first two chapters were supported by those in the third chapter, showing the success of the mapping model in a real-world example.

Regression Model

In the first chapter, I tested the mapping model against a multiple linear regression model, as the predominant model for quantitative judgment in the literature. Regression models have been widely and successfully used to capture judgment policies in many areas of social research (e.g., Hammond, 1996; Brehmer & Brehmer, 1988). However, researchers have questioned if people possess the cognitive capacities to perform the rather complex calculation required by a regression analysis. In this vein, it has been contended that regression models do not capture the process underlying a decision, even if they accurately predict decisions' outcomes (Hoffman, 1960; Gigerenzer & Todd, 1999; see also Doherty & Brehmer, 1997). However, many researchers have argued that this criticism does not necessarily affect the idea that judgments follow a linear additive estimation process. According to this argument participants assign each cue a subjective weight and then add the weighted cues (Einhorn et al., 1979; Brehmer, 1994; Juslin et al., 2003; Juslin et al., in press). Multiple linear regression analysis is employed to estimate the subjective weights that participants assigned to the cues, but without the claim that it reflects the process of how participants determined the weights.[1]

Consistent with this literature supporting the success of linear additive models to describe human judgment (Juslin et al., in press; Juslin et al., 2003; Brehmer, 1994; Einhorn

[1] However, it should be noted that the assumption that the estimation process in fact follows a linear additive combination of the cue information makes it reasonable to impose restrictions on the regression model, for example, see the constraints assumed by the cue abstraction module of the *Sigma* model (Juslin et al., in press).

et al., 1979; Kalish, Lewandowsky, & Kruschke, 2004; Anderson, 1981; Hammond, 1996), the results in Chapter 1 (Study 3) showed that a linear regression model described participants' estimations well, if the criterion was a linear additive function of the cues.

Interestingly, in this task, the correct cue weights could be abstracted easily, which could have enhanced the reliance on a linear additive strategy. Due to the deterministic nature of the task, the correct cue weights could be estimated from any two training objects, differing only on this cue during the training phase (Juslin, et al., in press). This was not possible in the tasks with skewed criteria or a nonlinear relation between the cues and the criteria, which could be one of the cues indicating to the participants that a linear additive strategy can not be successfully applied in these tasks.

Exemplar Model

In Chapters 1 and 2, I tested the mapping model against an exemplar-based model (Juslin et al., 2003). Exemplar models have been quite successful in describing behavior in categorization (Nosofsky, 1986; Kruschke, 1992), and were recently successfully put forward as a model for quantitative estimation (Juslin et al., in press). One advantage of exemplar models is that they can offer an accurate solution to tasks that can not be successfully solved by rule-based processes (Juslin et al. in press; Olsson, Enkvist, & Juslin, 2006). In this vein, Juslin and colleagues argued for a shift to exemplar-based processing if the criterion is a nonlinear function of the cues, and thus a linear additive model, such as multiple linear regression, could not successfully predict the criterion. However, the results in my studies suggested that these claims need to be further specified. The exemplar model was the best model to describe participants' estimations if participants had no prior knowledge about the cues and could not easily acquire knowledge during the training phase. Nevertheless, if these requirements were met, I found, consistent with Juslin and colleagues (in press), support for a spontaneous shift to exemplar-based processing. This is notable, as Olsson et al. (2006) recently had only found a shift to exemplar-based strategy if participants were explicitly instructed to use this strategy. Olsson and colleagues argued that a shift only occurs spontaneously if the reliance on exemplar knowledge promised accurate performance at the beginning of the training phase. As in Study 2 (Chapter 2), the exemplar model led to an accurate performance during training, which could have increased the accessibility of exemplar-based strategies.

Overall, the results suggested that exemplar models in fact offer a valid description of human estimation processes, but that the situations in which they are applied are rather specific. More precisely, people seem to only fall back on exemplar-based strategies if other rule-based models are not suited to solve the task.

A second point worth noting concerns the parameterization of the exemplar model. Throughout my dissertation, I considered two versions of the exemplar model: a complex version with a free parameter for every cue and a simplified version assuming that all cues are weighted equally. In the majority of the tasks, the simplified exemplar model was better in predicting participants' estimations than the more complex standard version, indicating that the original version of the exemplar model is prone to overfitting, and that the attention parameters need to be interpreted with caution. However, in the reanalysis of Experiment 1 by Juslin et al. (in press) and the second study in Chapter 2, there was a stable minority of participants best described by the complex exemplar model. This suggests that sometimes the additional parameters can prove to be necessary and informative, especially if not all cues are predictive and thus not used for the estimation. Likewise, Rehder and Hoffman (2005a, 2005b) showed that the parameter of exemplar models (see also Kruschke, 1992; Nosofsky, 1986) match the actual attention that participants allocate to the cues. However, it leaves the question of, how many parameters should be assumed a priori, and under which conditions the parameters can be reliably interpreted as reflecting the attention given to the cue (Medin & Schaffer, 1978; Juslin et al., 2003)? Results by Rehder and Hoffman (2005b) indicated that this could also be dependent on the duration of training, as they found a learning pattern where initially all cues were considered, but attention gradually concentrated on the relevant features.

QuickEst

In the first chapter, I also considered the heuristic QuickEst as a competitor for the mapping model and a possible alternative model for quantitative estimation (Hertwig et al., 1999). In particular, in the J-shaped condition, a good performance of QuickEst would have been expected. However, the overall support for QuickEst was rather weak. Though in the very first study about 20% of the participants were best described by QuickEst, it did not perform well in the second study. Likewise, an empirical study by Hausmann et al. (2007) did not find any support for QuickEst as a model of human behavior. However, in the studies presented here, one reason for the lack of support for QuickEst could lie in the design of the

task. Participants were presented with all relevant information simultaneously and free of charge. Research on inferential decision making in pair comparisons, however, indicated that noncompensatory strategies, like QuickEst, could be favored under time pressure or if information search is costly (Rieskamp & Hoffrage, in press; Payne, Bettman, & Johnson, 1988; Bröder & Schiffer, 2003). Like "tally," the mapping model is an information intensive strategy including all cues that are considered relevant. QuickEst, on the other hand, ignores a large part of the information, and thus QuickEst might be more readily employed if information has to be retrieved from memory or involves others' costs (Bröder & Schiffer, 2003).

Implications for the Process of Estimation

Assumptions of the Mapping Model

The mapping model makes explicit assumptions about the processes underlying estimation. Thus, the mapping model's success in describing the estimation process has specific implications regarding human estimation processes. According to the mapping model, people group objects together into a few categories based on the amount of evidence provided by the cues, and then select a typical estimate for each category, reflecting the central tendency of the objects falling into this category. More specifically, it implies that all cues are weighted equally, and that similar estimates follow from grouping objects together at an abstract level and not by similar configurations of cues. In these assumptions, the mapping model differs from the other models of estimation.

With the assumptions that all relevant cues are weighted equally, the mapping model differs from the regression model and the standard version of the exemplar model, which both propose that cues are weighted according to their importance. Overall, the results of my dissertation provided evidence for the equal weight approach assumed by the mapping model in nonlinear environments, but not for linear environments. In nonlinear estimation tasks, the mapping model accurately predicted participants' estimations in the laboratory. Likewise, the mapping model outperformed the regression model in a naturalistic estimation task. Furthermore, the exemplar model with a single attention parameter for all cues performed better than an exemplar model that potentially ignored cues, resonating with research highlighting the good performance of unit weight models in prediction tasks (Dawes, 1979). Although the unit weight approach of the mapping model was successful even if the cues

differed clearly in their predictiveness (Chapter 1), the results in Chapter 2 indicated that this might be limited to situations where knowledge about the cues is available. When the cues differed substantially in their predictiveness and participants had no prior knowledge, at least some participants were better captured by an exemplar model, allowing for differential weighting of the cues. This is also consistent with results by Rehder and Hoffman (2005b) who reported a shift of attention measured by eye movements to relevant cues during training.

A second related assumption of the mapping model is that object category membership is computed on the level of the summed evidence ignoring which specific cue contributed to the cue sum. Here, the mapping model differs from the exemplar model. While the exemplar model determines the similarity of two objects by the number of matches on the cues, and thus puts emphasis on specific configurations of cues, the mapping model assumes that objects are grouped together based on the total evidence provided by the cues, regardless of how many cues actually match. The clear qualitative pattern in the studies in Chapter 2 provided evidence that, at least if knowledge about the cue directions is available, participants' behavior corresponded to the assumptions of the mapping model, estimating similar values for objects with the same number of positive cues regardless of the dimension the cues were in. This also relates to Brunswik's idea of vicarious functioning (1952), that is, cues can be replaced by each other to reach the same judgment.

Adaptive Behavior in Quantitative Estimation

A further clear result from the experimental studies was that no single model could explain participants' behavior in all situations. Instead, the results indicated that whatever model was most successful in describing participants' estimations was a function of the environment. Similar to pair comparison tasks, participants adaptively shifted their estimation strategies to match the structure of the task (Todd & Gigerenzer, 2007; Rieskamp & Otto, 2006; Rieskamp, Busemeyer, & Laine, 2003; Payne et al. 1993; Juslin et al., in press). Consistent with the simulation study in Chapter 1, the mapping model performed better than or as good as the other models if the criterion followed a skewed or linear distribution. Likewise, the mapping model or the exemplar model was best if the criterion was a nonlinear function of the cues and thus a linear regression model was not suited to solve the task. However, if the criterion was a linear function of the cues, and thus linear regression was the optimal model to solve the task, participants' estimations were

consistently best described by a regression model. Moreover, the last paper indicated that the adaptive match between models and task structures was not just a result of the artificial nature of the task environment. In a real-world estimation task with a skewed criterion, when predicting sentence magnitude for low-level crimes, the mapping model also outperformed the regression model.

Likewise, the shift from the mapping model to exemplar-based processing, as reported in Chapter 2, can be regarded as adaptive. As knowledge about the correct cue directions is indispensable for the accurate performance of the mapping model, the ease with which the mapping model can be applied is closely tied to prior knowledge about the cues. If the cue directions are clear, the mapping model only demands minimal computation and can be correctly executed with little training. Thus, relying on the mapping model leads to a computational advantage if it can be applied to master the estimation task from the beginning. However, if the mapping model was not easily applicable because detecting the correct cue directions was difficult and a linear additive model could not successfully be applied, shifting to an exemplar-based estimation process can be considered an adaptive response to the task.

Though my results indicate an adaptive shift in processing dependent on the task structure, it remains unclear if the shift is due to an automatic error-driven learning process (e.g., Rieskamp & Otto, 2006; Erickson & Kruschke, 1998; Ashby, Alfonso-Reese, Turken, & Waldron, 1998) or to deliberate and voluntary processing (Haider, Frensch, & Joram, 2005). Although I did not model the learning process, it seems reasonable to assume that participants did not commit, at the onset of the task, for one type of processing, but learnt during the task which type of processing was most successful (Rehder & Hoffman, 2005b). However, it seems probable that also deliberative and controlled processes are involved. Recently, Haider and colleagues (2005) suggested that explicit knowledge stems from voluntary inferential processes. This suggests that learning the cue directions and thus the application of the mapping model, if no prior knowledge is available, could depend on a voluntary inferential effort of the participants to acquire this information.

Model Selection Methods

The goal of my dissertation was to determine which model was best suited to describe the participants' estimations. However, this raises methodological concerns, especially if models of differing complexity are compared, as just selecting the best-fitting model often

leads to the wrong choice (Robert & Pashler, 2000; Pitt, Myung, & Zhang, 2002). Although flexible models, that is, models with more free parameters, are better able to fit a specific dataset, they run the risk of overfitting the data. That is, they not only capture the systematic variance due to the underlying process but also fit random variance in the data. Thus, the best-fitting models are not necessarily best to accurately predict new data, making it indispensable to take model complexity into account for model selection. In my research, I addressed the problem of model selection with different methodologies.

Generalization Method: Out of Sample Prediction

In Chapters 1 and 2, I implemented a generalization test (Busemeyer & Wang, 2000). In a first step, I set the models' free parameters to equate the models' flexibilities by fitting them to a training set. Next, I predicted participants' estimates for a test set by computing the test objects' criterion values based on the obtained parameter values. The test set consisted of "old" exemplars, that is, test objects with the same cue values as the training objects, and of "new" objects, that is, test objects that they did not encounter during training, forcing the models to make sample predictions. Generalization tests go beyond pure cross validation: They not only ensure that only a model capturing the process underlying the estimation process is able to make accurate predictions, they also warrant that good model performance is not restricted to the objects encountered during training, but can be generalized out of the tested sample. However, they make the underlying assumption that the same processes govern the generation of data in both samples.

Qualitative Tests

Although quantitative measures of model performance are informative and allow a first test, if a model can capture human behavior, they do not offer any insights if the model assumptions actually correspond to the cognitive process generating the data. Furthermore, models often make very similar predictions, making it difficult to differentiate between them on a pure quantitative level. In the first chapter, I addressed this problem by selecting test objects on which the models differed in their predictions to increase the possibility to differentiate between the models. In the second chapter, I went one step further and tested qualitative predictions to underpin the quantitative model test. Qualitative tests are highly desirable because they can be constructed to be widely independent of model parameters, and they allow a better test of the models' assumptions (Pitt, Kim, Navarro, & Myung,

2006). My goal was to provide some evidence that the participants' behavior actually corresponds to the model assumptions about the estimation process. For this I focused on differences in the core assumptions the models make about the estimation process. For one the mapping model assumes that objects with a same cue sum are grouped together and receive the same criterion value as estimate. However, objects with differing cue sums are assigned to different categories and thus receive differing estimates. In contrast, the exemplar model relies on the similarity relations of the test objects to the training objects, which can be similar for two objects with differing cue sums. However, if two objects are maximally different, that is they do not match on a single cue, it is probable that they will also differ in their similarity relations to the training objects and thus in the estimates for the criterion. Based on these model assumptions, I constructed test conditions in which the models differed in their ordinal predictions, largely independent from the model parameters. Thus, when the participants' estimations matched the model predictions, this gave a strong indication that the model in fact captured the cognitive process underlying the estimations.

Bayesian Model Averaging

In the third paper, I relied on a different methodology for model selection. Similar to Chapters 1 and 2, one methodological problem was that I was comparing models of differing complexity, and thus different abilities to fit a data set. However, this study offered a further methodological problem as the relevant cues for the estimation task were not clear, but one goal of the analysis was to identify which predictors reliably influenced sentencing. In regression analysis, often methods based on significance testing, such as stepwise regression procedures, are used to find the best model to describe the data and to quantify the impact that the cues have on the estimation. However, these methods are often unreliable, potentially leading to different results if cues are stepwise included or excluded. Furthermore, focusing on a single model ignores the uncertainty involved in model selection and the conclusions that can be drawn from a single model. To address these problems, I chose a Bayesian Model Averaging method (BMA) by Raftery (1995). Based on the BIC approximation for the Bayes' factor (Schwartz, 1978; Raftery, 1995), the BMA method can be used to more reliably identify which models most probably underlie the data, and takes model complexity into account by penalizing a model for its number of free parameters. Moreover, it takes model uncertainty fully into account to determine which predictors have a significant impact on the estimation. Thus, it seemed to be a more reliable methodology to analyze the data.

Limitations and Extension of the Mapping Model

Though the mapping model was quite successful in describing participants' estimations, there are limits to its applicability. In the following, I will sketch some of the boundary conditions for the mapping model and how it could possibly be extended in the future.

Cue Selection

The mapping model does not have a mechanism incorporated to decide which cues should be included or to stop search for further information, but works on the assumption that all relevant cues are included in the estimation. This makes the assumption that prior knowledge about which cues are important is available and can be incorporated into the analysis, or the cues are preselected for their relevance (Brehmer, 1994). Thus, in real-world estimation tasks, the applicability of the mapping model could be limited because often an enormous amount of possible relevant cues can be identified, and knowledge about the cues' quality is not easily available (Brehmer & Brehmer, 1988). One way to solve this problem is to employ statistical methods, such as the BMA (Raftery, 1995), to identify which cues influenced the estimation processes.

However, a second possibility would be to implement a search and a stopping rule to model how the decision to include or exclude a cue is made (Gigerenzer & Todd, 1999).

Cue Weighting

The mapping model assumes that all cues are weighted equally; an assumption which was largely supported by the data in my dissertation. In a similar vein, unit weight linear models have been found to be as good or better as proper regression models, providing evidence for the robustness of a unit weight approach (Dawes, 1979; Einhorn & Hogarth, 1975). In this vein Dawes and Corrigan (1974) wrote:
"The whole trick is to decide what variables to look at and then know how to add." (p. 105).

However, the assumption of equal weights is a simplification which will not hold for all situations. It has been repeatedly shown that, in tasks with few cues, participants are able to differentially weight cues and can learn to ignore irrelevant cues (e.g., Castellan, 1973; Brehmer, 1973; Klayman, 1988; Kruschke & Johansen, 1999). Furthermore, Rehder and Hoffman (2005b) showed that spatial attention measured by eye movements was eventually

restricted to relevant cues. If several predictive cues are available, it seems unrealistic that the correct weights for all cues are correctly learnt, making a unit weight approach more probable. However, if only two or three cues are available and furthermore strongly differ in their validity, it seems probable that humans would learn not to rely on all cues, but concentrate their attention on the cues offering predictive information. Thus, in this situation, the simplification by the mapping model might lead to worse performance and not reflect the behavior of the participants, restricting it to situations with several predictive cues.

Extrapolation

Similar to the exemplar model, the mapping model does not extrapolate over the range of criterion values encountered during training. This seems to be a reasonable assumption if multiple cues are available and the environment is nonlinear (Juslin et al, 2003; Juslin et al., in press). Likewise, the results of Study 3 (Chapter 1) only provided evidence for extrapolation if the criterion was a linear additive function of the cues. However, research has shown that people are able to extrapolate in a one-dimensional function learning task (e.g., DeLosh, Busemeyer, & McDaniel, 1997; Kalish et al., 2004). Moreover, in the second study (Chapter 2), participants extrapolated over the experienced range in a condition in which the participants' estimations were otherwise well-described by the mapping model. This suggests that it could be plausible to consider an extrapolation mechanism for the mapping model. For instance, if the maximum is known and an object falling outside of the existing categories is encountered, a new category could be formed, and a typical criterion value falling between the criterion for the closest category and the maximum value estimated.

Continuous Cue Information

In the first two papers, I presented a version of the mapping model relying on binary cues. However, cues often provide more finely-graded information which can be used for the estimation. It seems to be sensible to assume that the continuous information is used and not just reduced to binary information. Thus, in the third chapter, I extended the mapping model to apply it to continuous cues. This extension differed in some respect from the binary version in Chapters 1 and 2, even though the model predictions are equivalent. More specifically, the continuous version of the mapping model assumes that, consistent with range frequency theory (Parducci, 1974), the perception of the magnitude of cues' values is

normalized, a mechanism which was not necessary in the binary version. Second, in the continuous version of the mapping models, the cues are integrated by averaging the cue values instead of adding positive cue information. Furthermore, in the first two papers, the number of categories and category membership is determined by the number of positive cues, but in the continuous version, the number of categories is set to seven (Miller, 1956), which are found by dividing the range of averaged cue values into equally sized categories.

As the continuous version of the mapping model made identical predictions in the tasks reported in Chapters 1 and 2, it was impossible to evaluate the two versions against each other in this work. However, from a theoretical perspective, relying on an averaging approach seems plausible, resonating with research on information integration (Anderson, 1965, 1967; Juslin et al., in press). Similarly, the range frequency theory provides a psychological plausible theory of how continuous cues are perceived (Parducci, 1974). Moreover, the results from Chapter 3 can be seen as a first support that the mapping model can be successfully extended to continuous cues. However, a more rigorous examination with an experimental approach of a continuous version is certainly necessary.

Generalizability and Applications of the Mapping Model

The success of the mapping model in predicting sentencing recommendations indicated that it can serve as a model of quantitative estimation in real-world tasks; in particular if similar conditions, as identified in the first two chapters, are encountered. This could be more frequent than appears at first glance: In many estimation tasks, we encounter in our daily lives that we possess explicit knowledge about the task, in particular about the cues. Knowledge about the task can not only be acquired through personal experience but can also be socially transmitted. For example, legal or medical education consists, to a large part, of transmitting knowledge about which cues are predictive in a specific task, such as diagnosing a specific disease or deciding if a defendant violated the law. Furthermore, skewed distributions are frequent. Because general growth processes commonly generate power law distributions (Gabaix, 1999), diverse phenomena ranging from city sizes to record sales or the size of computer files follow J-shaped distributions (for a review, see Schroeder, 1991). Thus, the conditions for the successful application of the mapping model could frequently be at hand. In a similar vein, consistent with the general idea of the mapping model, research in diverse areas of psychology has shown that unit weight summary indices of risk and protective factors are often the most reliable predictor to assess the risk of

juvenile delinquency or childrens' intelligence scores (e.g., Sameroff, Seifer, Baldwin, & Baldwin, 1993). In sum, the applicability of the mapping model is not restricted to laboratory or legal decision-making tasks, but can easily be employed to model quantitative estimations in a variety of areas.

Conclusion

Past research on quantitative estimation has almost exclusively relied on linear regression models to model the human estimation processes. In spite of the success of regression models in predicting the outcome of estimation, these models have been criticized for not capturing the cognitive process underlying estimations (Gigerenzer & Todd, 1999; Hoffman, 1960). Recently, alternative computational models were proposed for the area of quantitative judgment and estimation (e.g., Juslin et al., 2003; in press). My dissertation presents an important contribution to this literature, proposing a new cognitively inspired theory for quantitative estimation that outperformed current models of estimation in capturing peoples' behavior. In particular, in situations in which linear regression did not capture human behavior, the mapping model offered a plausible alternative solution. Furthermore, the mapping model explained peoples' estimations not only in several laboratory studies but also in a real-world environment. This suggests that its applicability is not restricted to laboratory tasks, but can be potentially employed in a diverse set of tasks. Thus, the mapping model offers an interesting extension for the adaptive toolbox (Gigerenzer & Todd, 1999), providing a further tool for quantitative estimations.

A second contribution of my research concerns the link between environmental structures and cognitive processing, extending existing research on the adaptive nature of human decision making to quantitative estimation (Payne et al., 1993; Gigerenzer, Todd, & the ABC Research Group, 1999). More specifically, my work showed that specific task structures differentially affected cognitive components that were essential for the models' assumptions about the estimation task. Consequently, model performance was, to a high degree, a function of the environment. In sum, my research not only highlights the impact of the environment on cognitive processing but also the importance of precise assumption about cognitive processes for psychological research.

References

Albers, W. (2001). Prominence theory as a tool to model boundedly rational decisions. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality: The adaptive toolbox* (pp. 297–317). Cambridge, MA: MIT Press.

Albrecht, H. J. (1980). Strafzumessung und Vollstreckung bei Geldstrafen unter Berücksichtigung des Tagessatzsystems [Sentencing and enforcement in fining under consideration of the daily payment system]. Berlin: Duncker & Humblot.

Anderson, N. H. (1965). Averaging versus adding as a stimulus combination rule in impression formation. *Journal of Experimental Psychology, 70,* 394–400.

Anderson, N. H. (1967). Averaging model analysis of set-size effect in impression formation. *Journal of Experimental Psychology, 75,* 158–165.

Anderson, N. H. (1981). Foundations of information integration theory. New York: Academic Press.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149-178.

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica, 87,* 137–154.

Brehmer, B. (1973). Single-cue probability learning as a function of the sign and magnitude of the correlation between cue and criterion. *Organizational Behavior and Human Decision Processes, 9*, 377-395.

Brehmer, A., & Brehmer, B. (1988).What have we learnt about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 75–114). Amsterdam: Elsevier/North Holland.

Brehmer B., & Joyce, C. R. B. (Eds.). (1988). *Human judgment: The SJT view.* Amsterdam: Elsevier/North Holland.

Bröder, A. (2000). Assessing the empirical validity of the take-the-best heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1332–1346.

Bröder, A., & Schiffer, S. (2003). Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology: General, 132*, 277–293.

Brown, N. (2002). Real world estimation: Estimation modes and seeding effects. In B. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 321–359). San Diego, CA: Academic Press.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*, 511–534.

Brown, N. R., & Siegler, R. S. (1996). Long-term benefits of seeding the knowledge base. *Psychonomic Bulletin and Review, 3*, 385–388.

Bruns, H-J. (1985). *Das Recht der Strafzumessung, Eine systematische Darstellung für die Praxis* (2. Aufl.) [*The law on sentencing: A systematic account for legal practitioners* (2nd ed.]. Köln: Heymanns.

Bruns, H-J. (1988). Die Bedeutung des Durchschnitts-, Regel- und des Normalfalles im Strafzumessungsrecht. Mögliche Orientierungspunkte für die Eingliederung der Tat

und des Strafmaßes in die Stufenfolge des Rahmens? [The relevance of the average, the norm, and the normal case in criminal sentencing. Possible points of reference for the placement of the offense and the sentence in the levels of the sentencing range] *Juritische Zeitung, 43*, 1053-1058.

Brunswik. E. (1952). *Conceptual framework of psychology*. Chicago: University of Chicago Press.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Busemeyer, J. R., & Wang, Y-M. (2000). Model comparisons and model selection based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*, 171–189.

Castellan, N. J. (1973). Multiple-cue probability learning with irrelevant cues. *Organizational Behavior and Human Decision Processes, 9*, 16-29.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal, 23*, 41–65.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods and applications*. San Diego: Academic Press.

Colwell, L. H. (2005). Cognitive heuristics in the context of legal decision making. *American Journal of Forensic Psychology, 23*, 17-41.

Conrad, F. G., Brown, N. R., & Cashman, E. R. (1998). Strategies for estimating behavioural frequency in survey interviews. *Memory, 6,* 339–366.

Davis, T. L., Severy, L. J., Kraus, S. J., & Whitaker, J. M. (1993). Predictors of sentencing decisions: The beliefs, personality variables, and demographic factors of juvenile justice. *Journal of Applied Social Psychology, 23,* 451-476.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 968–986.

Dhami, M., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making, 14,* 141–168.

Dhami, M. K. (2003). Psychological models of professional decision making. *Psychological Science, 14*, 175-180.

Doherty, M., & Brehmer, B. (1997).The paramorphic representation of clinical judgment: A thirty-year retrospective. In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections and controversies* (pp. 537–551). Cambridge: Cambridge University Press.

Doherty, M. E., & Kurz, E. (1996). Social judgement theory. *Thinking and Reasoning, 2*, 109–140.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory process model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Ebbesen, E. B., & Konecni, V. J. (1975). Decision making and information integration in the courts: The setting of bail. *Journal of Personality and Social Psychology*, *32*, 805–821.

Ebbesen, E. B., & Konečni, V. J. (1981). The process of sentencing adult felons. In B. D. Sales (Ed.), *The trial process* (pp. 413-458). New York: Plenum.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13*, 171–192.

Einhorn, J. H., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Regression models and process tracing analysis. *Psychological Review, 86,* 465–485.

Engel, C., & Gigerenzer, G. (2006). Law and heuristics: An interdisciplinary venture. In C. Engel & G. Gigerenzer (Eds.), *Heuristics and the law* (pp. 1-16). Cambridge, MA: MIT Press.

Engen, R. L., & Gainey, R. R. (2000). Modeling the effects of legally relevant and extralegal factors under sentencing guidelines: The rules have changed. *Criminology, 38*, 1207-1229.

Englich, B., & Mussweiler, T. (2001). Sentencing under uncertainty: Anchoring effects in the courtroom. *Journal of Applied Social Psychology, 31*, 1535-1551.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*, 188-199.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107-140.

Fishbein, M., & Ajzen, I. (1980). *Understanding attitudes and predicting social behavior*. New York: Prentice Hall.

ForsterLee, R., ForsterLee, L., Horowitz, I. A., & King, E. (2006). The effect of defendant race, victim race, and juror gender on evidence processing in a murder trial. *Behavioral Sciences and the Law, 24*, 179-198.

Foth, E. (1985). Strafschärfung/Strafmilderung—eine noch unerledigte Frage der Strafzumessung [Aggravating and mitigating sentences—outstanding issues for sentencing]. *Juristische Rundschau*, *10*, 397-399.

Friedman, W. J. (1993). Memory for the time of past events. *Psychological Bulletin, 113*, 44–66.

Friedman, W. J. (2004). Time in autobiographical memory. *Social Cognition, 22*, 591–605.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics*, *114*, 739–767.

Gigerenzer, G., & Kurz, E. (2001). Vicarious functioning reconsidered: A fast and frugal lens model. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 342–347). New York: Oxford University Press.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 3–34). New York: Oxford University Press.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*, 592–596.

Gigerenzer, G. (2006). Heuristics. In G. Gigerenzer & C. Engel (Eds.), *Heuristics and the law* (pp. 17-41). Cambridge, MA: MIT Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Gonzalez-Vallejo, C., & Bonham, A. (in press). Aligning confidence with accuracy: Revisiting the role of feedback. *Acta Psychologica*.

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2001). Inside the judicial mind: Heuristics and biases. *Cornell Law Review, 86*, 777–830.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition, 65*, 197-230.

Haider, H., Frensch, P. A., & Joram, D. (2005). Are strategy shifts caused by data-driven processes or by voluntary processes? *Consciousness and Cognition, 14*, 495–519.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62,* 255–262.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Hammond, K. R. & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Harries, P. A., & Harries, C. (2001). Studying clinical reasoning. Part 2: Applying social judgment theory. *British Journal of Occupational Therapy, 64*, 285–292.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining inference and prediction*. New York: Springer.

Hausmann, D., Läge, D., Pohl, R., & Bröder, A. (in press). Testing the QuickEst: No evidence for the Quick-Estimation heuristic. *European Journal of Cognitive Psychology*.

Helson, H. (1964). *Adaptation-level theory*. New York: Harper & Row.

Helversen, B. von, & Rieskamp, J. (in press). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*.

Henning, K., & Feder, L. (2005). Criminal prosecution of domestic violence offenses: An investigation of factors predictive of court outcomes. *Criminal Justice and Behavior, 32*, 612-642.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer, P.M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 209–234). New York: Oxford University Press.

Hertwig, R. (2006). Do legal rules rule behavior. In C. Engel & G. Gigerenzer (Eds.), *Heuristics and the law* (pp. 391-411). Cambridge, MA: MIT Press.

Hertwig, R., Hoffrage, U., & Sparr, R. (2007). The QuickEst heuristic: How it benefits from an imbalanced world. *Manuscript in preparation*.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57,* 116–131.

Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 301-315.

Hogarth, R. M., Gibbs, B. R., McKenzie, C. R. M., & Marquis, M. A. (1991). Learning from feedback: Exactingness and incentives. *Journal of Experimental Psychology: Learning Memory and Cognition, 17*, 734–752.

Johnson, B. D. (2006). The multilevel context of criminal sentencing: Integrating judge- and county-level influences. *Criminology, 44*, 259-297.

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003a). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29*, 924–941.

Juslin, P., Karlsson, L., & Olsson, H. (in press). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*.

Juslin, P., Olsson, H., & Olsson, A-C. (2003b). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132*, 133–156.

Juslin, P., & Persson, M. (2002). PROBabilities from Exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26*, 563–607.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Kahneman D., & Tversky A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582–591.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review, 111*, 1072–1099.

Karlsson, L., Juslin, P., & Olsson, H. (2004). Representational shifts in a multiple-cue judgment task with continuous cues. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 648–653). Mahwah, NJ: Cognitive Science Society.

Kautt, P., & Spohn, C. (2002). Crack-ing down on black drug offenders? Testing for interactions among offenders' race, drug type, and sentencing strategy in federal drug sentences. *Justice Quarterly, 19*, 1-35.

Kautt, P.M. (2002). Location, location, location: Interdistrict and intercircuit variation in sentencing outcomes for federal drug-trafficking offenses. *Justice Quarterly, 19*, 633–671.

Klayman, J. (1988a). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 317-330.

Klayman, J. (1988b). On the how and why (not) of learning from outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 115–162): Amsterdam: North Holland.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. Journal of Experimental Psychology. *Learning, Memory, and Cognition, 25*, 1083–1119.

Langer, W. (1994). *Staatsanwälte und Richter, Justitielles Entscheidungsverhalten zwischen Sachzwang und lokaler Justizkultur [Prosecutors and judges, legal decision making between necessity and local legal culture]*. Stuttgart: Enke.

Leiser, D., & Pachman, O. (2007). On the complexity of traffic judges' decisions. *Manuscript submitted for publication.*

Levy, M., & Solomon, S. (1997). New evidence for the power-law distribution of wealth. *Physica, 242*, 90–94.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Meier, B-D. (2001). *Strafrechtliche Sanktionen* (2.Aufl.) [*Criminal sanctions* (2nd ed.)]. Berlin: Springer.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 27, *775-799*.

Mösl, A. (1981). Zum Strafzumessungsrecht [On sentencing law]. *Neue Zeitung für Strafrecht*, 131-135.

Mösl, A. (1983). Zum Strafzumessungsrecht [On sentencing law], *Neue Zeitung für Strafrecht*, 160-164.

Myung, J. I., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 422–437), London: SAGE Publications.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 53–65.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of William K. Estes* (Vol. 1, pp. 149–167). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7*, 375–402.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79.

Olsson, A. C., Enqvist, T., & Juslin, P. (2006). Go with the flow! How to master a nonlinear multiple-cue judgment tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1371–1384.

Olsson, H., Wennerholm, P., & Lyxzén, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning , Memory, and Cognition, 30*, 936–941.

Ojmarrh, M. (2005). A meta-analysis of race and sentencing research: Explaining the inconsistencies. *Journal of Quantitative Criminology, 21*, 439-466.

Oswald, M. E. (1994). *Psychologie des richterlichen Strafens [Psychology of judicial sentencing]*. Stuttgart: Enke.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407–418.

Parducci, A. (1974). Context effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (vol. 2). New York: Academic Press.

Patalano, A. L., Smith, E. E., Jonides, J., & Koeppe, R. A. (2001). PET evidence for multiple strategies of categorization. *Cognitive, Affective & Behavioral Neuroscience, 1,* 360-370.

Payne, J. W., Bettman, J. R., & Johnson E. J. (1993). *The adaptive decision maker.* New York: Cambridge University Press.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 534-525.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113*, 57–83.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163.

Raftery, A. E., Madigan, D., & Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*, 179-191.

Rehder, B., & Hoffman, A. B. (2005a). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 811–829.

Rehder, B., & Hoffman, A. B. (2005b). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51*, 1–41.

Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 1371–1384.

Rieskamp, J., Busemeyer, J. R., & Laine, T. (2003). How do people learn to allocate resources? Comparing two learning theories. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *29*, 1066–1081.

Rieskamp, J., & Hoffrage, U. (in press). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*.

Rieskamp, J., & Otto, E. P. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236.

Robert, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Ruback, R. B., & Wroblewski, J. (2001). The federal sentencing guidelines: Psychological and policy reasons for simplification. *Psychology, Public Policy and Law, 7,* 739-775.

Sameroff, A. J., Seifer, R., Baldwin, A., & Baldwin C. (1993). Stability of intelligence from preschool to adolescence: The influence of social and family risk factors. *Child Development, 64*, 80–97.

Schäfer, G. (2001). *Praxis der Strafzumessung* (2. Aufl.) [*The practice of sentencing* (2nd ed.)]. München: Beck.

Schroeder, M. (1991). *Fractals, chaos, power laws: Minutes from an infinite paradise*. New York: Freeman.

Schünemann, B. (1988). Daten und Hypothesen zum Rollenspiel zwischen Richter und Staatsanwalt bei der Strafzumessung [Data and hypotheses on the role play between judges and prosecutors in sentencing]. In G. Kaiser, H. Kury, & H. J. Albrecht (Eds.), *Kriminologische Forschung in den 80er Jahren. Projektberichte aus der BRD* (pp. 265-280). Freiburg: Max Planck Institute for Foreign and International Criminal Law.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*, 3–21.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

Tata, C. (1997). Conceptions and representations of the sentencing decision process. *Journal of Law and Society, 24*, 395–420.

Tata, C. (1998). The application of judicial intelligence and 'rules' to systems supporting discretionary judicial decision-making. *Artificial Intelligence and Law, 6*, 203-230.

Theune, W. (1985a). Grundsätze und Einzelfragen der Strafzumessung; aus der Rechtsprechung des Bundesgerichtshofs (Teil 1) [Principles and individual issues of sentencing: From the jurisdiction of the Federal Court of Justice in Germany (part 1)]. *Strafverteidiger*, *4*, 162-168.

Theune, W. (1985b). Grundsätze und Einzelfragen der Strafzumessung; aus der Rechtsprechung des Bundesgerichtshofs (Teil 2) [Principles and individual issues of sentencing: From the jurisdiction of the Federal Court of Justice in Germany (part 2)]. *Strafverteidiger*, *5*, 205–210.

Todd, P. M., & Gigerenzer, G. (2007) Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science, 16*, 167–171.

Tröndle, H., & Fischer, T. (2007). *Strafgesetzbuch und Nebengesetze* (54. Aufl) [*Penal code and comments* (54th ed.)]. München: Beck.

Van Duyne, P. (1987). Simple decision making. In D. C. Pennington & S. Lloyd-Bostock (Eds.), *The psychology of sentencing: Approaches to consistency and disparity* (pp. 143-158). Oxford: Centre for Socio-Legal Studies.

Wigton, R. S. (1996). Social judgement theory and medical judgement. *Thinking and Reasoning, 2*, 175–190.

Wryobeck, J. M., & Rosenberg, H. (2005). The association of client characteristics and acceptance of harm reduction: A policy capturing study of psychologists. *Addiction Research and Therapy, 13*, 461–476.

Zedeck, S., & Kafry, D. (1977). Capturing rater policies for processing evaluation data. *Organizational Behaviour and Human Performance, 18*, 269–294

List of Tables

List of Figures

Erklärung

Hiermit bestätige ich die Ersteinreichung der vorgelegten Arbeit als Dissertation. Ich versichere, dass ich diese Arbeit eigenständig und nur mit Hilfe der genannten Quellen erstellt habe. Im Weiteren erkläre ich, dass ich die Promotionsordnung der Humboldt-Universität zu Berlin zur Kenntnis genommen habe.